The international journal of science / 30 April 2020

nature

Index
Annual Tables

MOLECULAR SYRINGE

High-resolution structures shed light on the workings of an antibacterial nanomachine

The international journal of science / 30 April 2020

nature

COVID-19 digital apps need due diligence

Governments see coronavirus apps as key to releasing lockdowns. But they must be effective and the data secure.

In the toolkit of strategies to stop the spread of SARS-CoV-2, more countries are reaching for smartphone apps. When phones with such an app are close together, they exchange information — in some cases creating a log of who a phone's owner has been near. These 'contacts' will be alerted if they have been close to an infected person. Such apps can complement a country's overall COVID-19 control strategies — including testing, contact tracing, isolation and social distancing — but they cannot serve as a replacement for them, or the thousands of contact-tracing teams they require.

Like any health-care intervention, coronavirus apps need to conform to the highest standards of safety and efficacy. But despite the pandemic's global nature, countries are developing apps independently, and there are no global standards — which is rightly raising concerns.

Some countries are already starting to use phones to record data, including names, addresses, gender, age, location, disease symptoms and COVID-19 test results. For example, users of Australia's COVIDSafe app, launched last weekend, will be contacted by health officials if an app user they have had close contact with tests positive for COVID-19. Germany's app, which is still in development, will also use actual test results. Australia is storing data centrally, but, after much debate, and expressions of concern from researchers, Germany's app will store coronavirus data on individuals' phones. Egypt's app, launched earlier this month, uses a phone's location services to alert users if they have been near anyone with COVID-19.

Use of all of these apps is voluntary, as it should be. In most cases, the apps are being developed by governments working with technology companies and researchers. But, considering that citizens are being asked to give up their personal data, there has been little national public consultation. Another cause for concern is the fact that there is scant published evidence on how effective these apps will be at either identifying infected people who have not been tested or, if widely used, stopping the spread of the disease. Governments are excitedly pointing out the benefits, but are saying less about the risks.

Key questions need answers

One serious concern is accuracy. Apps that link to official validated tests are more likely to give accurate results. An alert based on self-diagnosis that turns out to be wrong — a false positive — could, of course, be corrected. But if incorrect information has been sent to a large group of contacts, it will have caused unnecessary alarm, and could have wrongly sent people into isolation for weeks.

> **"Apps should not be rolled out without pilot studies or risk assessments being published."**

An equally important concern is privacy. As we have pointed out before, it is becoming easier to identify individuals from anonymized data sets. Researchers have shown that it is possible to re-identify individuals even when anonymized and aggregated data sets are incomplete (L. Rocher *et al. Nature Commun.* **10**, 3069; 2019).

Researchers are also raising concerns about the decision some countries have taken to store data centrally. Earlier this month, nearly 300 researchers signed an open letter reminding governments that data stored on individual phones are more secure, and that data stored centrally are more susceptible to hacking.

COVID-19 apps have, to some extent, been inspired by the experiences of South Korea and Singapore. South Korea, in particular, is regarded as a model because it avoided severe lockdowns. Some 3 months after the outbreak spread to the country, only a handful of new cases are being reported daily and 244 deaths have been recorded in total.

But the foundation for South Korea's COVID-19 response is a comprehensive testing strategy, backed by a nationwide network of contact-tracers who interview infected people and trace their contacts. The strategy includes the use of phone alerts, but not the type of phone app being developed elsewhere. More importantly, it is based on a degree of surveillance that people in many other countries would find hard to accept.

When a person tests positive for COVID-19, a text alert is sent to everyone living nearby. The alert typically includes a link to a detailed log of the infected person's movements — in some cases to the nearest minute — which are reconstructed from public data, such as closed-circuit television cameras. But the government is also permitted to access confidential records, such as credit-card transactions. The data are then stored centrally by government agencies.

Much attention has also been paid to Singapore's app, which now has more than one million users — roughly one-fifth of the population. But it still means that in any encounter between two randomly chosen people, there is only a 4% chance that both will have the app. This points to one of the deepest flaws in digital contact-tracing plans anywhere: the fact that only a fraction of any population is likely to have the app at all. And such efforts will miss out anyone who, for any reason, doesn't have a smartphone.

It's not that digital contact tracing shouldn't be done, but it should not be a substitute for human contact-tracing teams; nor should it be seen as a replacement for necessary COVID-19 testing. And apps should not be rolled out without pilot studies or risk assessments being published.

Speed is, of course, of the essence — but so is due diligence and due process. This includes public dialogue; more involvement from researchers, including those who study ethics, law and public engagement; and a cast-iron commitment from governments that the information being harvested is secure and will only ever be used for the reasons it is being requested.

nature

Education must fix its data deficit

An extensive assessment of education shows that only 61% of children worldwide will complete secondary education.

How can we ensure that every child gets a good education? Achieving quality schooling for all is one of the United Nations' Sustainable Development Goals (SDGs). But one of the largest studies of its kind confirms that countries are some way from ensuring a school place for about 260 million children currently not attending school.

On page 636, a team led by Emmanuela Gakidou, who studies health metrics at the University of Washington in Seattle, reports an analysis of education data from 195 countries and territories for the period from 1970 to 2018 (J. Friedman *et al. Nature* **580**, 636–639; 2020). From this, the authors built a model to forecast what the picture would look like in 2030 – the year by which countries agreed to meet all 17 SDGs and their associated targets.

Although almost 90% of children are expected to be completing primary school by 2030, only 61% of young adults (aged 25–29) will have finished secondary education. (Moreover, the modelling was done before the COVID-19 pandemic, which is preventing many children from attending school.) If all of the targets for the education SDG are to be met, all young people need to complete both primary and secondary education – and this should be free of charge.

Although the education SDG might not be met, the overall trend has been in the right direction. In 1970, just 50% of 25–29-year-olds had completed primary school. That had risen to 83% by 2018, and is projected to reach 89% by 2030.

Similarly, in 1970, only about 20% of the 25–29 age cohort had finished secondary school. And in the countries of North Africa and the Middle East, just 7% had reached this milestone. By 2030, the researchers predict, around 75% of the same age cohort in these countries will have completed secondary schooling.

Attainment of tertiary education – defined as 15 or more years of education – has also improved around the world, with some of the largest gains again occurring in North Africa and the Middle East. But even in the world's best-performing regions, the team predicts that just half of young people will be completing tertiary education by 2030.

What needs to be done so that more children can receive a full education? A more concerted push is required, especially from decision makers, and this push needs to be backed by more-precise data that build on existing knowledge. We already know, for example, that on a global level, girls are less likely than boys to complete their schooling – although this gender gap has nearly closed, which will enable more countries to make progress in the SDG for gender equality. We also know that children living in some rural areas are less likely than those in urban regions to progress to secondary education, or even finish primary school.

> "Education is essential to solving the many problems the world is facing."

At the same time, we know which groups of children are less likely to enter school at all, or to progress from primary to higher levels of education. In general, it is those whose household incomes are low, which sometimes means children have to work to contribute to the household budget. Other children affected include those with disabilities, those in regions experiencing conflict and those belonging to minority groups. Disadvantage in any form tends to hold children back. Furthermore, in many countries, children are not at school because there are not enough publicly funded schools of sufficient quality.

But we also need data on these factors at the level of villages, towns, cities and districts, and they need to be tracked systematically so that progress can be monitored, especially in low-income countries. Such data will shine a light on which groups need the most help, allowing educational authorities – and funders – to target their efforts.

Better data must also be collected on educational achievements, known as learning outcomes. Such data are patchy and difficult to compare across countries, because there is no agreed international standard. Although in some countries, children completing primary school are expected to be able to read to a set standard, in many parts of the world they are not. The World Bank is working with the UN children's fund, UNICEF, to create standardized assessments for learning outcomes and – along with other organizations – is calling for these to be part of the drive to achieve the SDG for education.

How data have made a difference

Data have already helped to achieve the gains we are seeing in primary education, but these gains came out of a strong desire to effect change. Twenty years ago, 83% of children were enrolled in primary education. UN member states considered this unacceptable. They collected relevant national and local data, and discovered that the reasons for low attendance included malnutrition and poor provision for children in rural areas. Decision makers responded with measures such as mobile schools, and free – or subsidized – school meals. By 2015, primary-school enrolment had reached 91%.

Such an approach can also be used to grow secondary and tertiary education. Researchers can help nations to understand more about who is not attending school and why – and then put that information into the hands of decision makers and funders.

Education is a right – as enshrined in the Universal Declaration of Human Rights and in the UN Convention on the Rights of the Child. Education is also linked to better health and prosperity. It is essential to solving the many problems the world is facing, from climate change to the coronavirus pandemic. The generation of children now in schools is going to inherit a world that seems to grow more unstable with each passing decade. We need to equip them with every tool they will need – and that includes the best education.

World view

By John Nkengasong

Let Africa into the market for COVID-19 diagnostics

Africa is boosting its capacity to respond to COVID-19, but lack of solidarity will cost lives, warns Africa CDC head John Nkengasong.

The first case of COVID-19 in Africa was reported in Egypt on 14 February 2020. Since then, 52 countries in Africa have reported more than 30,000 cases and about 1,400 deaths from the new coronavirus. This count is likely to be an underestimate; Ethiopia has run about 11,000 tests — only 10 for every 100,000 people. Much-richer South Africa has run about 280 per 100,000. For Australia, the number is about 2,000; for the United States, 1,560.

First the good news. African countries are used to widespread testing for pathogens such as HIV, tuberculosis and malaria. This expertise can easily be adapted for SARS-CoV-2 testing. The Africa Centres for Disease Control and Prevention (Africa CDC), which I lead, held the first of its training sessions in early February. By mid-March, 43 countries had gained competence to test for the virus — if appropriate reagents were accessible.

But they are not. The collapse of global cooperation and a failure of international solidarity have shoved Africa out of the diagnostics market. With its lack of hospitals and high prevalence of conditions such as HIV, tuberculosis, malaria and malnutrition, Africa could see COVID-19 mortality rates higher than elsewhere, even in children. It will be higher still the more slowly we implement testing. No country can securely eliminate COVID-19 — or its devastating economic domino effects — if the disease becomes rampant across a continent of 1.3 billion people. For Africa to get ahead of the pandemic, we need to scale up testing fast.

Lack of access to diagnostics is Africa's Achilles heel. When SARS-CoV-2 was first reported, genome sequences were made available within weeks and several groups in Asia and Europe started producing in-house tests. Africa lacked this capacity and had to wait for the tests to be introduced, a tardy 'trickle-down' of diagnostics. The situation has now become worse: a race is on by the powerful to acquire whatever COVID-19 tests are available.

This is not a question of demanding charity. African countries have funds to pay for reagents but cannot buy them. To solve this problem, we need solidarity both across the world and within the continent. But instead of global solidarity, global protectionism has prevailed, with more than 70 countries imposing restrictions on the export of medical materials. Wealthier countries should reserve some fraction of these supplies for export; that would cut their own risk that the disease will be reintroduced. Where export markets are open, African countries must band together to negotiate as one large customer, rather than as many small buyers fighting for a seat at the table.

> "African countries have funds to pay for reagents but cannot buy them."

John Nkengasong is the director of the Africa Centres for Disease Control and Prevention in Addis Ababa, Ethiopia.
e-mail: nkengasongj@africa-union.org

Cooperation across Africa is starting to happen. Africa CDC has a plan to distribute one million test kits by mid-May, which we started implementing earlier this month.

The strategy to increase testing for COVID-19 is fourfold. First, we need to pool the procurement and distribution of tests across the continent. This will create synergies and block counterproductive competition. Second, we have to work with non-government laboratories and the private sector to roll out testing on the subnational level. In many countries, samples must be shipped to a centralized diagnostic laboratory, which adds cost and delay. Third, we need to make sure our testing technologies can use the existing platforms that have been the backbone for large-scale testing for HIV and tuberculosis.

Finally, we need to speed up the production of test kits within Africa. Plans for production are under way in Kenya, Morocco, Senegal and South Africa. In April, Africa CDC launched an initiative called Partnership to Accelerate COVID-19 Testing (PACT). The goal is to reach 10 million tests in the next four months, although this timescale falls far short of serving our very real immediate needs.

Africa CDC has developed guidelines for targeted testing that include every person with pneumonia, and people in clusters of disease that could be COVID-19. PACT will focus on coordinating efforts to act on test results so that countries can isolate infected individuals, perform contact tracing and quarantine exposed people (ideally, in their own homes). This means mobilizing hundreds of community workers and building partnerships between local government officials, public-health experts, social scientists and other community leaders. To do this we plan to draw on the experience of the African Health Volunteers Corps in fighting Ebola in West Africa and the Democratic Republic of the Congo. Training of community workers must cover ethics, confidentiality, privacy, discrimination, stigmatization and individual rights.

For PACT to succeed and save lives, it needs international support and partnership with the private sector, not interference, and support from the African Union COVID-19 Response Fund. Government leaders, finance and health ministers and global-health experts must be willing to work together for a collective win, even when leaders feel pressure from their own nations. The decision of the finance ministers of the G20 group of nations, the World Bank Group and regional development banks to rapidly implement a US$200-billion emergency response package, and to temporarily suspend debt payments for the poorest countries, is welcome and essential.

Hundreds of thousands of Africans shouldn't need to die because of a global crisis that requires global action and global solidarity. If Africa loses, the world loses.

JOHN NKENGASONG

News in brief

THOUSANDS VOLUNTEER FOR CONTROVERSIAL VACCINE STUDY

Momentum is building to speed the development of coronavirus vaccines by intentionally infecting healthy volunteers with the virus. A grass-roots effort has attracted about 3,900 potential volunteers for the controversial approach, known as a human-challenge trial.

The effort, called 1Day Sooner, is not affiliated with groups or companies developing or funding coronavirus vaccines. But co-founder Josh Morrison hopes to show that there is broad support for human-challenge trials.

Typical vaccine trials take a long time because thousands of people receive either a vaccine or a placebo, and researchers track who becomes infected in the course of their daily lives. A challenge study could, in theory, be much faster: a smaller group of volunteers would receive a candidate vaccine and then be intentionally infected with the virus, to judge the efficacy of the immunization.

"We want to recruit as many people as possible who want to do this, and pre-qualify them as likely to be able to participate in challenge trials should they occur," says Morrison.

Wellcome, a biomedical-research funder in London, has begun discussing the ethics and logistics of a human-challenge trial for a coronavirus vaccine, says Charlie Weller, head of its vaccines programme. But she says it is unclear whether such a trial could actually speed up vaccine development.

EPIC ARCTIC RESEARCH MISSION FORCED TO BREAK OUT OF ICE

When scientists were planning MOSAiC — a pioneering expedition that would trap a research vessel in Arctic sea ice for one year — they considered the North Pole's hazards. But no one anticipated a pandemic.

The travel restrictions and flight cancellations caused by the coronavirus outbreak have now forced mission planners to take a tough decision. *Polarstern*, the German ship central to the expedition, will temporarily leave its position in the ice to exchange its crew, and will abandon the research camp where it has been frozen since last October.

The disruption is a blow to the mission's researchers, who have created a unique platform from which to study climate change in the Arctic. Although they hope to refreeze the ship at the same camp, the interruption will leave a hefty gap in the data set.

Scientists plan to leave the research camp mostly intact, although certain measurements must stop. But there are some autonomous instruments that will continue to function, taking measurements of, for example, wind speed, temperature, pressure and humidity.

"We're going to do the best we can with these constraints," says Matthew Shupe, an atmospheric and oceanic scientist at the University of Colorado Boulder and co-leader of MOSAiC. "But in the end, it's a bummer."

MORE US LABS COULD BE PROVIDING TESTS FOR CORONAVIRUS

A survey of more than 4,000 researchers in the United States suggests that better coordination at an institutional and national level could make hundreds of thousands more tests for coronavirus available.

To find out what is preventing molecular-biology laboratories with the capabilities to run tests from doing so, Giovanni Paternostro, a biomedical researcher at Sanford Burnham Prebys Medical Discovery Institute in La Jolla, California, and Joshua Graff Zivin, an economist at the University of California, San Diego, sent a survey to 35,000 principal investigators who had received grants from the US National Institutes of Health (NIH) in 2018.

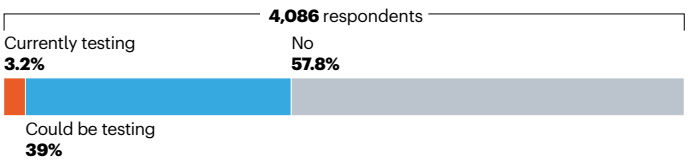
Of the more than 4,000 researchers who responded within the first week, about 130 were already running tests to detect the new coronavirus (see 'Keen to help'). Nearly 1,600 said that they had the main tool needed to run tests, a real-time PCR machine, and operated under the biosafety conditions required for working with pathogenic organisms such as coronavirus. But they were not testing.

Both groups — those who are testing and those who could — were asked what they would need to process more tests or to begin testing. Resources such as reagents and funding were a popular response for both groups, as was coordination by the NIH or their own institution. About 95% of labs not currently testing said they needed more information on protocols and regulations, such as the key Clinical Laboratory Improvement Amendments (CLIA) certification for providing clinical-test results. But 43% of labs currently doing testing said that they could do with more information in these areas as well.

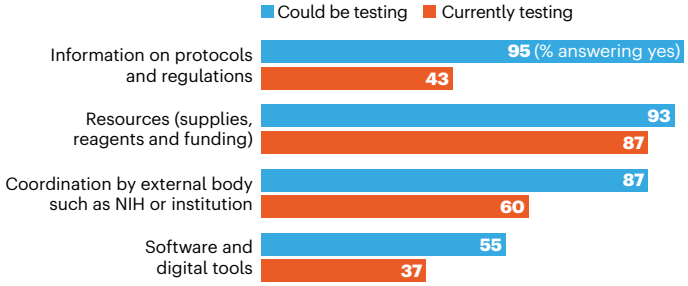
KEEN TO HELP

In a survey of almost 1,600 US laboratories, more than 40% met the basic necessary requirements to perform testing for the coronavirus that causes COVID-19. Many said that they could do better with more resources and with better coordination.

Is your laboratory providing COVID-19 testing — or does it have the basic necessary elements to do so?



Would addressing one or more of the following needs allow your lab to provide the test, or to provide it more efficiently if you are already doing it?



L TO R: MICHAEL GUTSCHE; SOURCE: DATA FROM G. PATERNOSTRO & J. G. ZIVIN

News in focus

GREG BAKER/AFP/GETTY

Antibody tests have been promoted as a way to get people back to work — but the reliability of their results is unknown.

WILL CORONAVIRUS ANTIBODY TESTS REALLY CHANGE EVERYTHING?

The rapidly developed tests have been touted as society's way out of widespread lockdowns, but scientists say it will be a while before they are as useful as hoped.

By Smriti Mallapaty

British Prime Minister Boris Johnson called them a "game changer". Antibody tests have captured the world's attention for their potential to help life return to normal by revealing who has been exposed, and might now be immune, to the new coronavirus.

Dozens of biotechnology companies and research laboratories have rushed to produce the blood tests. And governments around the world have bought millions of kits, in the hope that they could guide decisions on when to relax social-distancing measures and get people back to work. Some have even suggested that the tests could be used as an 'immunity passport', giving the owner clearance to interact with others again.

Many scientists share this enthusiasm. The immediate goal is a test that can tell healthcare and other essential workers whether they are still at risk of infection, says David Smith, a clinical virologist at the University of Western Australia in Perth. In the future, the tests could assess whether vaccine candidates give people immunity.

But as with most new technologies, there are signs that the promises of COVID-19 antibody tests have been oversold, and their challenges underestimated. Kits have flooded the market, but most aren't accurate enough to confirm whether an individual has been exposed to the virus.

And even if tests are reliable, they can't indicate whether someone is immune to re-infection, say scientists. It will be a while before kits are as useful as hoped, says Smith.

The UK government learnt about this the hard way after it ordered 3.5 million tests from several companies in late March, only to discover later that none of these tests

performed well enough.

"No test is better than a bad test," says Michael Busch, director of the Vitalant Research Institute in San Francisco, California.

Antibody tests are also being used by researchers globally to estimate the extent of coronavirus infections at a population level. These surveys test a portion of the population and use those results to estimate infections among the broader community. More than a dozen groups worldwide are doing such studies.

Flood of tests

When a virus invades the body, the immune system produces antibodies to fight it. Kits detect the presence of antibodies using components from the virus, known as antigens. Tests generally fall into one of two categories: lab tests that need to be processed by trained technicians and take about a day, and point-of-care tests that give rapid, on-the-spot results within 15 minutes to half an hour. Several companies, including Premier Biotech in the United States and China-based Autobio Diagnostics, offer point-of-care kits, which are designed to be used by health professionals to check whether an individual has had the virus – but some companies market them for people to use at home.

The tests don't detect the virus itself, so have limited use in diagnosing active infections, say health agencies. But in some countries, such as the United States and Australia, tests are being used in some cases to diagnose people who are suspected to have COVID-19, but who test negative on a standard PCR test, says Smith. (A study[1] by researchers at Shenzhen Third People's Hospital in China found that PCR tests did not always diagnose people infected with the virus.)

Early studies in those who have recovered from COVID-19 have detected three kinds of SARS-CoV-2-specific antibody, and manufacturers and research institutes have developed tests that target these antibodies. For instance, the German biopharmaceutical company EUROIMMUN has developed a lab test that detects SARS-CoV-2-specific immunoglobulin G and immunoglobulin A.

Because of the ongoing emergency, the US Food and Drug Administration (FDA) has relaxed the rules that govern the use of such tests. It has authorized their use in laboratories and by health-care workers to diagnose active COVID-19 infections, with the disclaimer that they have not been reviewed by the FDA and that results should not be used as the sole basis for confirming that someone has the disease. Australia has also introduced similar emergency authorizations.

These measures are appropriate given the pandemic situation, says Smith. Antibody tests in people who might be actively infected can be an important part of managing patients at hospitals, and can aid contact tracing, although the results need to be interpreted cautiously, he says.

One problem, however, is that most kits have not undergone rigorous testing to ensure they're reliable, says Busch.

To verify their accuracy, kits need to be trialled on large groups of people that include hundreds who have had COVID-19, and hundreds who haven't, says Peter Collignon, a physician and laboratory microbiologist at the Australian National University in Canberra. But so far, most test assessments have involved only some tens of individuals, because they have been developed quickly.

It seems that many tests available now are not accurate enough at identifying people who have had the disease, a property called test sensitivity, or at detecting those who haven't been infected, known as test specificity. A high-quality test should achieve 99% or more sensitivity and specificity, adds Collignon. That means that testing should turn up only about one false positive and one false negative for every 100 true positive and true negative results.

"No test is better than a bad test."

But some commercial antibody tests have recorded specificities as low as 40% early in the infection. In an analysis[2] of 9 commercial tests available in Denmark, 3 lab-based tests had sensitivities in the range of 67–93% and specificities of 93–100%. In the same study, 5 out of 6 point-of-care tests had sensitivities ranging from 80% to 93%, and 80–100% specificity, but some kits were tested on fewer than 30 people. Testing was suspended for one kit. Overall, the sensitivity of all the tests improved over time, with the highest sensitivity recorded two weeks after symptoms first appeared.

Point-of-care tests are even less reliable than tests being used in labs, adds Smith. This is because they use a smaller sample of blood – typically from a finger prick – and are conducted in a less controlled environment than a lab, which can affect their performance. They should be used with caution, he says. The World Health Organization recommends that point-of-care tests be used only for research.

Timing is crucial

One unknown that affects both kinds of test is the interplay between timing and accuracy. If a test is done too soon after a person is infected and the body hasn't had time to develop the antibodies the test is designed to detect, it could miss an infection. But scientists don't yet know enough about the timing of the body's immune responses to SARS-CoV-2 to say exactly when specific antibodies develop.

By contrast, false positives crop up if a test uses an antigen that doesn't only target antibodies produced to fight SARS-CoV-2, and instead picks up antibodies for another pathogen as well, says Smith. An analysis[3] of EUROIMMUN's antibody test found that, although it detected SARS-CoV-2 antibodies in three people with COVID-19, it returned a positive result for two people with another coronavirus.

Ironing out all these issues takes time and involves trial and error, says Collignon. It took several years to develop antibody tests for HIV with more than 99% specificity, he says.

Infection doesn't equal immunity

Another big question surrounding antibody tests is the extent to which being infected with a pathogen confers immunity to reinfection. To have protective immunity, the body needs to produce a certain type of antibody, called a neutralizing antibody, which prevents the virus from entering cells.

But it's not clear whether all people who have had COVID-19 develop these antibodies. An unpublished analysis[4] of 175 people in China who had recovered from COVID-19 and had mild symptoms reported that 10 individuals produced no detectable neutralizing antibodies. These people had been infected, but it's unclear whether they have protective immunity, says Wu Fan, a microbiologist at Fudan University in Shanghai, China, who led the study.

So far, researchers say they have not seen any evidence that people can get reinfected with the virus. "We should presume that once you have been infected, your chance of getting a second infection two to three months later is low," says Collignon. But how long that protective immunity will last is not known.

Even if it becomes clear that most people do develop neutralizing antibodies, most tests currently don't detect them.

The fact that most antibody tests can't detect neutralizing antibodies is also relevant because some politicians are pushing the use of these tests to clear those with past COVID-19 infections to interact with others again, with what is known as an immunity passport. Researchers are trying to determine whether the antibodies detected by current kits can act as a proxy for protective immunity, says Smith.

Despite the challenges, when reliable antibody tests are available, they could be important for understanding which groups of people have been infected and how to stop further spread, says Collignon. They could even be used to diagnose active infections when PCR tests fail, adds Smith.

1. Zhao, J. *et al. Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa344 (2020).
2. Lassaunière, R. *et al.* Preprint at medRxiv https://doi.org/10.1101/2020.04.09.20056325 (2020).
3. Okba, N. M. A. *et al. Emerg. Infect. Dis.* https://doi.org/10.3201/eid2607.200841 (2020).
4. Wu, F. *et al.* Preprint at medRxiv https://doi.org/10.1101/2020.03.30.20047365 (2020).

MANISH SWARUP/AP/SHUTTERSTOCK

Hydroxychloroquine is used to treat malaria and some autoimmune diseases.

CHLOROQUINE HYPE DERAILS CORONAVIRUS DRUG TRIALS

People hoping to use malaria drugs to fight COVID-19 are turning away from clinical trials of other therapies.

By Heidi Ledford

People with COVID-19 who arrive at the Salvador Zubirán National Institute of Medical Sciences and Nutrition in Mexico City to search for treatment can choose from a menu of clinical trials, carefully presented by a worker trained to offer an unbiased portrait of the potential risks and benefits.

But neurologist Sergio Iván Valdés-Ferrer already knows which trial most will choose — and it's not his. Instead, many people opt for one involving hydroxychloroquine, a malaria drug that has been touted by US President Donald Trump and other influential figures as an effective coronavirus treatment.

"There's a tremendous bias," says Valdés-Ferrer, who is studying the effects of a dementia drug on COVID-19. "Studies of any other drug that are enrolling all ages and degrees of severity are in big trouble."

Hydroxychloroquine and its close chemical cousin chloroquine have attracted disproportionate attention in the coronavirus pandemic, spurred by preliminary studies and added publicity from political leaders such as Trump and French President Emmanuel Macron. So far, there are very few data backing the idea that hydroxychloroquine works against coronavirus infection, yet the fervour surrounding it has created drug shortages and affected enrolment in clinical trials for other potential treatments.

Early evidence

The US Food and Drug Administration (FDA) has approved chloroquine and hydroxychloroquine to treat malaria and, because of the drugs' anti-inflammatory properties, to treat some autoimmune diseases, such as rheumatoid arthritis and lupus. In February, researchers showed that chloroquine could reduce coronavirus infection of human cells grown in the laboratory (M. Wang *et al. Cell Res.* **30**, 269–271; 2020). A few days later came a report of clinical trials involving people with COVID-19 in ten hospitals in China, which suggested that chloroquine treatment might shorten the duration of the disease (J. Gao *et al. Biosci. Trends* **14**, 72–73; 2020). Since then, a handful of small studies have been reported. None has shown definitively whether or not the drug can benefit people with COVID-19.

But the initial findings were sufficient for politicians keen to offer worried voters and suffering economies a glimmer of hope. Trump claimed that he had considered taking chloroquine as a precaution. Hospitals in Iran, New York, Spain and China turned to hydroxychloroquine and chloroquine as a standard therapy for people with COVID-19, despite guidance from the World Health Organization and several medical associations that the drugs should not be used for this purpose except in clinical trials.

The resulting race to take, or even to hoard, chloroquine led to global shortages, and there were reports of illness and deaths linked to overdoses in the United States and Nigeria.

Endangered studies

Some people don't want to participate in clinical trials that would require them to give up chloroquine treatments. This has made it difficult to enrol people in a trial of HIV drugs as potential COVID-19 treatments, says infectious-disease specialist Sung-Han Kim at the University of Ulsan College of Medicine in Seoul. Kim's story isn't unique. Psychiatrist Eric Lenze of Washington University in St. Louis, Missouri, recently launched a trial of an antidepressant that he hopes could lessen the immune response linked to some severe COVID-19 cases. The trial has so far enrolled just ten participants; three others declined to take part because they were already planning to take hydroxychloroquine.

In Iran, pathologist Alireza Ghaffarieh gave up plans to exclude chloroquine treatment from his trial of an iron-chelating medicine in people with COVID-19 at the Kermanshah University of Medical Sciences. Instead, he accepts participants who might be taking other medications, and hopes that this will not complicate interpretation of his results.

Delays in enrolment can endanger a clinical trial, particularly during a pandemic, says Prashant Malhotra, an infectious-disease specialist at North Shore University Hospital in Manhasset, New York. It's best if trials can be completed early, he says, before health-care systems become overwhelmed.

Researchers might have settled some of these issues weeks ago if there had been a rapid, international effort to develop a rigorous chloroquine clinical trial, says Ole Søgaard, an infectious-disease physician at Aarhus University Hospital in Denmark. Now, there are more than 100 clinical trials that aim to test chloroquine or hydroxychloroquine against COVID-19. It's a worthwhile effort, Søgaard says, despite the lack of evidence supporting the drugs. "Being able to cross something like hydroxychloroquine off the list and move on to other things would be a major achievement," he says. "Then you could shut down a lot of trials and replace them with something you believe in."

Additional reporting by Amy Maxmen.

FIRST MAJOR VIRTUAL PHYSICS MEETING SEES RECORD ATTENDANCE

The American Physical Society held its massive April Meeting online because of the coronavirus.

By Davide Castelvecchi

Despite some last-minute scrambling, the first major physics conference to be held in cyberspace was a success, according to many attendees.

The April Meeting of the American Physical Society (APS) was originally scheduled to take place on 18–21 April in Washington DC. But when the coronavirus pandemic made a physical gathering impossible, the organizers decided to hold the entire event online, and made registration free and open to everyone.

Whereas around 1,600–1,800 people typically attend the April Meeting, 7,267 registered this time, says Hunter Clemens, the APS director of meetings. And many participants say they were satisfied. "The virtual APS meeting has been by far the best online meeting I have attended," says Niels Warburton, an astrophysicist at University College Dublin.

In early March, the APS was one of the first large organizations outside of China — where the first outbreak of the virus was reported — to bear the brunt of the pandemic. The society decided to cancel its much larger March Meeting in Denver, Colorado, just 36 hours before it was due to start. Some of that meeting took place anyway: would-be attendees quickly organized unofficial versions of the scheduled sessions online.

Inspired in part by that surge of enthusiasm, the APS opted to hold its next major meeting online, rather than cancelling or postponing it. The society hired a company to provide the necessary online infrastructure and technological support. During the 4-day conference, it handled 175 live sessions, running up to 15 in parallel. The online platform for talks provided a chat window that appeared alongside the speaker's video, allowing attendees to exchange comments or links to relevant papers in real time. The APS also made an effort to recreate the social experience of a conference by organizing virtual meet-ups, and some delegates set up their own discussions using messaging tools such as Slack.

Although a virtual meeting is not the real thing, it was still a good idea given the circumstances, says Xiaochao Zheng, a nuclear and particle physicist at the University of Virginia in Charlottesville. "Many other conferences are cancelled, which are big disappointments for people who had planned to attend," she says. Lindley Winslow, an experimental physicist at the Massachusetts Institute of Technology in Cambridge, agrees. "In my field of neutrinos and dark matter, we have to do a lot of our meetings virtually. It works, but it is not as efficient as having everyone in the same room," she says. Still, she adds, because she had a newborn baby at home, "It was a bit of relief to not have to figure out how to travel."

The virtual meeting had some other advantages compared with a physical one. Live talks could be paused or rewound, a useful feature for those who missed details or wanted to spend more time pondering a crucial slide.

And watching talks from home eased a bit of the pressure of attending a large conference that would require dashing from one session to another across a vast convention centre. "I'm kinda loving the minimal FOMO [fear of missing out] when you're just feeling tired/introverted/overwhelmed that comes along with everyone being virtual," tweeted Claire Lee, a particle physicist at the Fermi National Accelerator Laboratory outside Chicago, Illinois.

The last-minute transition to cyberspace was not completely smooth, in part because it came long after the conference programme had been finalized. Although most speakers agreed to present their talks online, some did not. And some sessions, including many of the talks contributed by students, had to be pre-recorded to be watched 'on demand'. This created confusion among participants, some of whom found out too late that they had to upload their talk ahead of time. The APS is still allowing presenters to upload their videos after the meeting, Clemens says.

> **"The virtual APS meeting has been by far the best online meeting I have attended."**

Most attendees contacted by *Nature* found the conference useful. "My quick take-away is that it was more successful than I thought it was going to be," says Kelly Backes, a graduate student at Yale University in New Haven, Connecticut. "I got a lot more out of it than I expected."

Q&A

Sweden's coronavirus strategist

Sweden's relaxed approach to containing the coronavirus — largely using voluntary measures — has drawn sharp criticism, including from high-profile scientists. The nation's COVID-19 death rate is higher than that of its neighbours, which have imposed lockdowns, and it has seen large outbreaks in care homes. The strategy's architect is epidemiologist Anders Tegnell at Sweden's Public Health Agency, which advises the government. Tegnell spoke to *Nature* about the approach.

Can you explain Sweden's strategy?
As in many other countries, we aim to slow down the spread — otherwise the health-care system is at risk of collapse. This is not a disease that can be eradicated, at least until a working vaccine is made. The Swedish laws on communicable diseases are mostly based on individual responsibility. Quarantine can be contemplated for people or small areas, but we cannot lock down a region.

What evidence is the approach based on?
It is difficult to talk about the scientific basis of a strategy with this type of disease, because we do not know much about it and we are learning as we go. Lockdown, closing borders — nothing has a historical scientific basis, in my view. As a society, we are more into nudging: continuously reminding people to use measures. Closing down everything would be counterproductive.

The strategy has been criticized for being too relaxed. What is your response?
The public-health agency has released detailed regional modelling that comes to less-pessimistic conclusions than other researchers' models in terms of hospitalizations and deaths. We will see a lot more cases in the next few weeks, but that is just like any other country. I am confident schools will stay open. We underestimated the issues at care homes; we should have controlled this more thoroughly.

Interview by Marta Paterlini
This interview has been edited for length and clarity.

JONATHAN NACKSTRAND/AFP/GETTY

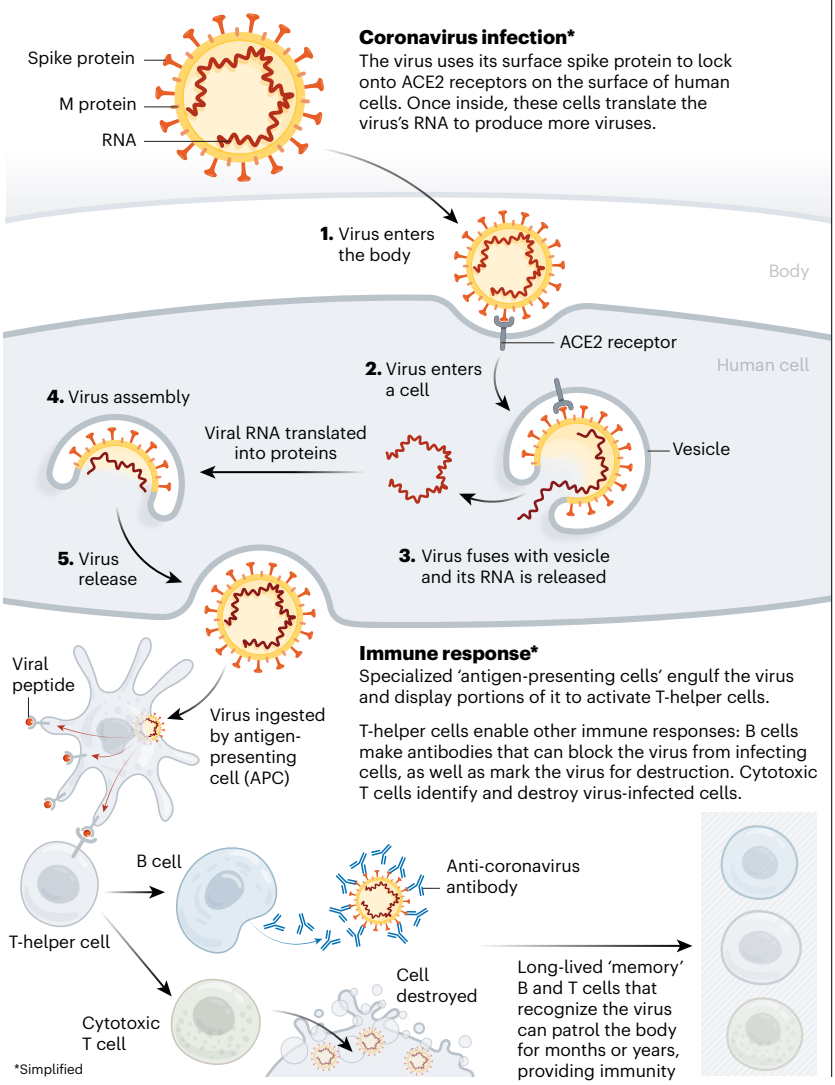
THE RACE FOR CORONAVIRUS VACCINES

By Ewen Callaway; design by Nik Spencer.

More than 90 vaccines are being developed against SARS-CoV-2 by research teams in companies and universities across the world. Researchers are trialling different technologies, some of which haven't been used in a licensed vaccine before. At least six groups have already begun injecting formulations into volunteers in safety trials; others have started testing in animals. *Nature*'s graphical guide explains each vaccine design.

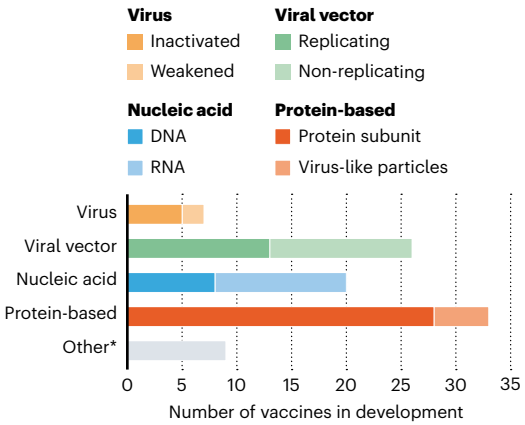
VACCINE BASICS: HOW WE DEVELOP IMMUNITY

The body's adaptive immune system can learn to recognize new, invading pathogens, such as the coronavirus SARS-CoV-2.



AN ARRAY OF VACCINES

All vaccines aim to expose the body to an antigen that won't cause disease, but will provoke an immune response that can block or kill the virus if a person becomes infected. There are at least eight types being tried against the coronavirus, and they rely on different viruses or viral parts.



* Other efforts include testing whether existing vaccines against poliovirus or tuberculosis could help to fight SARS-CoV-2 by eliciting a general immune response (rather than specific adaptive immunity), or whether certain immune cells could be genetically modified to target the virus.

VIRUS VACCINES

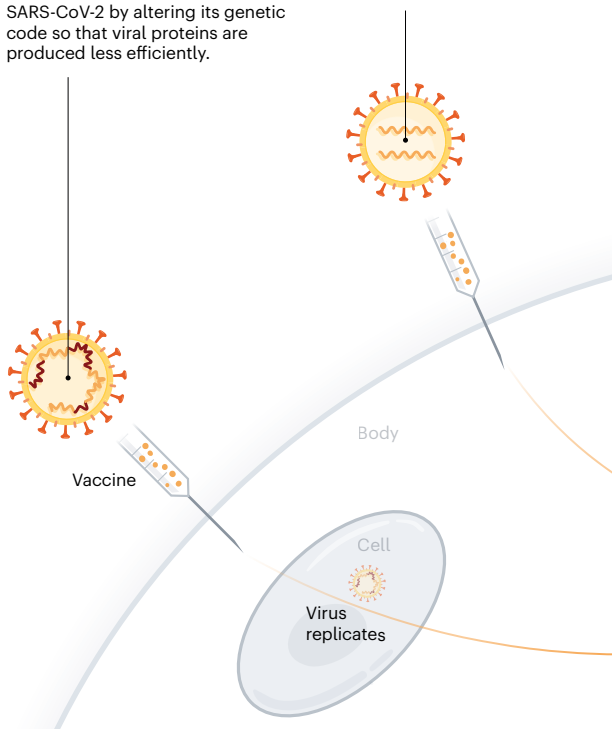
At least seven teams are developing vaccines using the virus itself, in a weakened or inactivated form. Many existing vaccines are made in this way, such as those against measles and polio, but they require extensive safety testing. Sinovac Biotech in Beijing has started to test an inactivated version of SARS-CoV-2 in humans.

Weakened virus

A virus is conventionally weakened for a vaccine by being passed through animal or human cells until it picks up mutations that make it less able to cause disease. Codagenix in Farmingdale, New York, is working with the Serum Institute of India, a vaccine manufacturer in Pune, to weaken SARS-CoV-2 by altering its genetic code so that viral proteins are produced less efficiently.

Inactivated virus

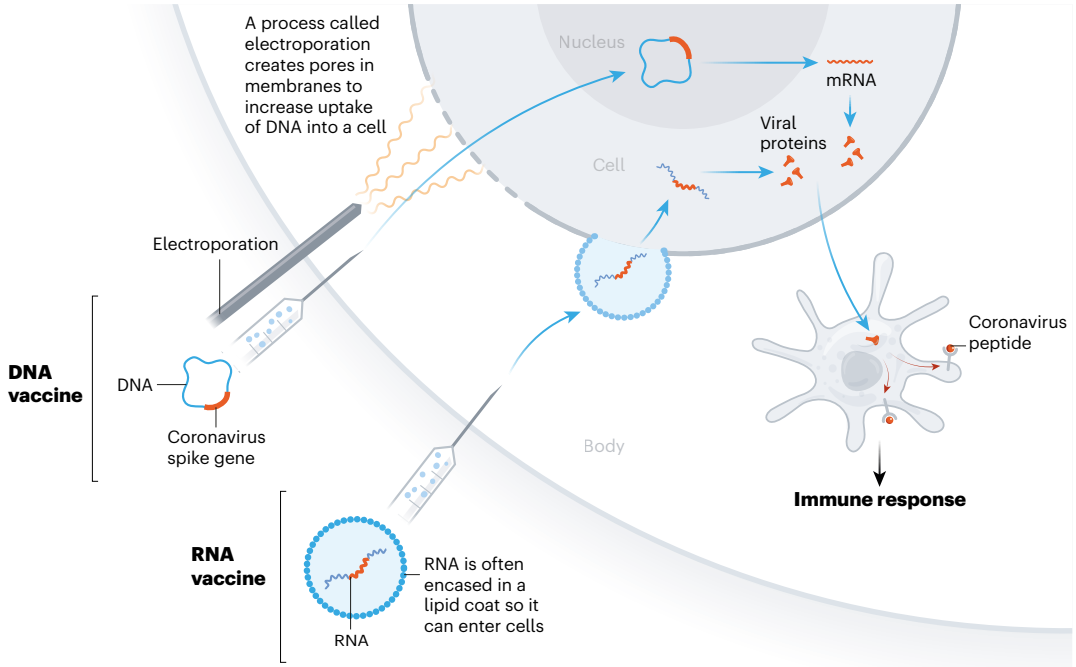
In these vaccines, the virus is rendered uninfectious using chemicals, such as formaldehyde, or heat. Making them, however, requires starting with large quantities of infectious virus.



NUCLEIC-ACID VACCINES

At least 20 teams are aiming to use genetic instructions (in the form of DNA or RNA) for a coronavirus protein that prompts an immune response. The nucleic acid is inserted into human cells, which then churn out copies of the virus protein; most of these vaccines encode the virus's spike protein.

RNA- and DNA-based vaccines are safe and easy to develop: to produce them involves making genetic material only, not the virus. But they are unproven: no licensed vaccines use this technology.



VIRAL-VECTOR VACCINES

Around 25 groups say they are working on viral-vector vaccines. A virus such as measles or adenovirus is genetically engineered so that it can produce coronavirus proteins in the body. These viruses are weakened so they cannot cause disease. There are two types: those that can still replicate within cells and those that cannot because key genes have been disabled.

Replicating viral vector (such as weakened measles)

The newly approved Ebola vaccine is an example of a viral-vector vaccine that replicates within cells. Such vaccines tend to be safe and provoke a strong immune response. Existing immunity to the vector could blunt the vaccine's effectiveness, however.

Non-replicating viral vector (such as adenovirus)

No licensed vaccines use this method, but they have a long history in gene therapy. Booster shots can be needed to induce long-lasting immunity. US-based drug giant Johnson & Johnson is working on this approach.

PROTEIN-BASED VACCINES

Many researchers want to inject coronavirus proteins directly into the body. Fragments of proteins or protein shells that mimic the coronavirus's outer coat can also be used.

Protein subunits

Twenty-eight teams are working on vaccines with viral protein subunits — most of them are focusing on the virus's spike protein or a key part of it called the receptor binding domain. Similar vaccines against the SARS virus protected monkeys against infection but haven't been tested in people. To work, these vaccines might require adjuvants — immune-stimulating molecules delivered alongside the vaccine — as well as multiple doses.

Virus-like particles

Empty virus shells mimic the coronavirus structure, but aren't infectious because they lack genetic material. Five teams are working on 'virus-like particle' (VLP) vaccines, which can trigger a strong immune response, but can be difficult to manufacture.

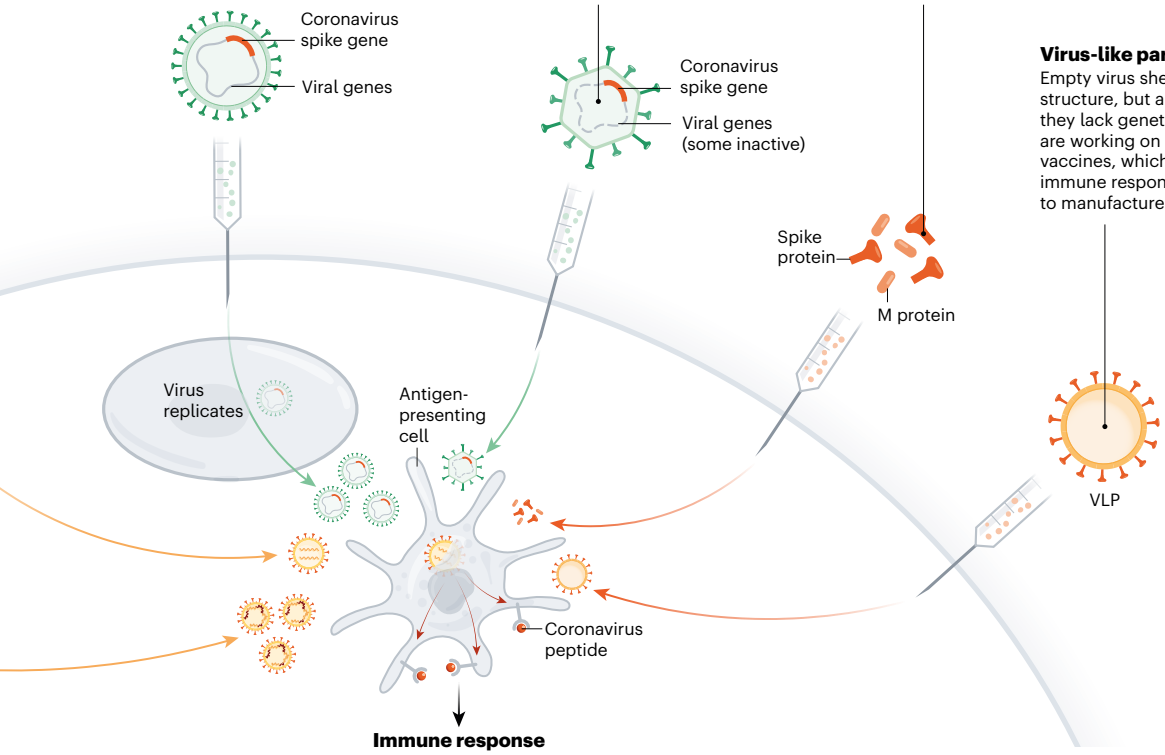


CHART SOURCES: *NATURE* ANALYSIS BASED ON WHO COVID-19 VACCINE LANDSCAPE/MILKEN INSTITUTE COVID-19 TREATMENT AND VACCINE TRACKER/T. THANH LE *ET AL. NATURE REV. DRUG. DISC.* HTTP://DOI.ORG/GGRNBR (2020)/F. AMANAT & F. KRAMMER *IMMUNITY* 52, 583–589 (2020)/W. SHANG *ET AL. NPJ VACCINES* 5, 18 (2020)

MARTIN STORZ/IMAGEBROKER/PICTURE ALLIANCE

Research facilities at CureVac in Tübingen, Germany, one of dozens of firms working on a coronavirus vaccine.

CAN THE WORLD MAKE ENOUGH CORONAVIRUS VACCINE?

Researchers warn production constraints and hoarding could limit SARS-CoV-2 vaccine supplies. **By Roxanne Khamsi**

As the world searches for a way to end the coronavirus pandemic, the race is on to find and produce a vaccine. Some optimistic forecasts suggest that one could be available in 12–18 months — but researchers are already warning that it might not be physically possible to make enough vaccine for everyone, and that rich countries might hoard supplies.

The production facilities needed will depend on which kind of vaccine turns out to work best. Some researchers say governments and private funders should give vaccine manufacturers money to ramp up their production capacity in advance, even if these facilities are never used. Although money has been pledged to help with this, the promises fall short of the billions of dollars that public-health experts say is needed.

Resources for coronavirus will also have to be balanced against the need for other vaccines. Manufacturing facilities around the world can churn out hundreds of millions of doses of influenza vaccine each year, and companies are used to stepping up production at times of high demand. But if billions of people need a new kind of vaccine for coronavirus, and firms continue making the normal array of shots against influenza, measles, mumps and rubella, and other diseases, there could be a production shortage, says David Heymann, an infectious-disease specialist at the London School of Hygiene and Tropical Medicine who heads a panel that advises the World Health Organization (WHO) on disease emergencies such as the COVID-19 pandemic.

The WHO says it is also working on a plan to ensure the equitable distribution of vaccines.

But how that could be enforced in practice isn't clear. "In a pandemic, the last thing we want is for vaccines to be exclusively accessed by countries that make them and not be universally available," says Mariana Mazzucato, an economist who heads the University College London Institute for Innovation and Public Purpose.

Supply constraints, both physical and political, are a "big worry", agrees Seth Berkley, who heads GAVI, the Vaccine Alliance — a public–private non-profit organization based in Geneva, Switzerland, that aims to increase access to immunizations around the world.

Pick a winner?

One big challenge in creating a lot of vaccine quickly is scaling up manufacturing, because the infrastructure needed will differ depending on the vaccine type.

The vaccine might consist of a weakened or inactivated version of the coronavirus, or some part of a surface protein or a sequence of RNA or DNA, injected into the body inside a nanoparticle or another virus, such as measles. It might need to be grown in vats of cells, created using a machine that synthesizes RNA or DNA, or even grown in tobacco plants.

If vaccines built from inactivated forms of SARS-CoV-2 prove most effective, it should be easier to estimate what it would take to churn out doses, because this industrial technology has been around since at least the 1950s, says Felipe Tapia, who studies bioprocess engineering at the Max Planck Institute for Dynamics of Complex Technical Systems in Magdeburg, Germany. That said, the production and purification of whole SARS-CoV-2 virus at high concentrations could require facilities with biosafety level 3 certification. These are scarce, Tapia says, and could be why very few companies say they are trying this approach.

At least a dozen companies are chasing the idea of injecting into the body formulations of RNA or DNA that would provoke our cells into making one of the proteins used by SARS-CoV-2. "RNA and DNA platforms may involve a simpler process — which is likely to make them easier to scale up," says Charlie Weller, head of the vaccines programme at Wellcome, a London-based biomedical research funder. But no vaccine with this approach has yet been approved for any disease in humans.

Moderna, headquartered in Cambridge, Massachusetts, which injected its first test RNA-based coronavirus vaccine into a volunteer in mid-March, is one firm trialling this plan: another is CureVac in Tübingen, Germany, which says it has the facilities necessary to produce up to 400 million doses a year of its RNA-based vaccine. Both efforts have received money from the Coalition for Epidemic Preparedness Innovations (CEPI), a fund based in Oslo that was launched in 2017 as a global alliance to finance and coordinate vaccines for outbreaks.

SANJIT DAS/BLOOMBERG/GETTY

Vials of vaccine (not for coronavirus) at a manufacturing plant in Pune, India.

CEPI has also announced funding for six other vaccine research teams, including a collaboration that wants to re-engineer a measles vaccine so that it produces a immunizing SARS-CoV-2 protein in the body. If that works, says Marie-Paule Kieny, a virologist and director of research at INSERM, France's national biomedical research agency in Paris, it's possible that measles vaccine-manufacturing facilities could be used to make a COVID-19 vaccine, but she cautions that it's likely that capacity would have to be increased so as not to disrupt the original focus.

This is a challenge that must be urgently and collectively addressed."

Other elements in the manufacturing process might create bottlenecks. 'Subunit' vaccines, which are composed of a SARS-CoV-2 protein, or a key fragment of one, often need an adjuvant — molecules added to boost the immune response. These might require ingredients that could become scarce during a pandemic, such as specific lipids, says Jaap Venema, chief science officer of US Pharmacopeia (USP), a non-governmental organization in Rockville, Maryland, that helps to set drug-quality standards.

Another idea to grow vaccines quickly is using plants. Cigarette giant British American Tobacco (BAT) said in April that it aims to grow vaccines (being developed by its subsidiary Kentucky BioProcessing) in fast-growing tobacco plants. But Venema says such plant-based vaccine products have extra regulatory hurdles to clear, including complying with rules for genetically modified organisms — which could make it very hard to fast-track the process.

Cash in advance

An open question is how to ensure that the world's governments and companies invest enough money now, so that vaccines can be made quickly in 2021. CEPI says that global funding of at least US$2 billion is needed to help develop candidate vaccines and manufacture them for trials, of which it has been promised $690 million by national governments. A further $1 billion is needed to manufacture and distribute a successful SARS-CoV-2 vaccine for the world, CEPI says. But many more billions of dollars might need to be plunged into helping companies scale up manufacturing capacity, even if that isn't ultimately used, CEPI chief executive Richard Hatchett told STAT.

Billionaire philanthropist Bill Gates, who co-chairs the Bill & Melinda Gates Foundation in Seattle, Washington, also says facilities should be built in advance. He told US media that his foundation would help to pay for this approach, "just so that we don't waste time" until we know which vaccine platform will be the most successful. But the Gates Foundation did not provide further details when contacted for this story.

One firm to secure a large investment is drug giant Johnson & Johnson, which in March announced a $1-billion partnership with the US government's Biomedical Advanced Research and Development Authority to develop a vaccine based on an engineered version of an adenovirus. This includes a plan to rapidly scale up capacity, with the goal of "providing global supply of more than one billion doses of a vaccine". (In an early indication on pricing, Paul Stoffels, the company's chief scientific

MEHDI CHEBIL/POLARIS/EYEVINE

A researcher at the Institut Pasteur in Paris works on re-engineering a measles vaccine to trigger an immune response to SARS-Cov-2.

officer, has suggested that this vaccine might theoretically cost around $10 or €10 per dose.)

Governments could help vaccine makers to plan ahead, says Ohid Yaqub, a health-policy researcher at the University of Sussex in Brighton, UK, by signalling how much vaccine they plan to purchase and who they would recommend to be immunized.

A step further would be to set up what are known as advanced market commitments to purchase drugs at a specific price ahead of the vaccine being approved, as has happened for the distribution for pneumococcal vaccine to children through GAVI.

Berkley and others also say that donor countries could sell bonds to investors as a way to finance vaccines for populations that cannot afford them. This approach has also been used successfully before: the International Finance Facility for Immunisation (IFFIm) to raise money for vaccines that GAVI has provided to children.

Laws against hoarding?

But even if lots of vaccine is made, there seems to be no way to force countries to share it. During the 2009 H1N1 influenza pandemic, Australia was among the first to manufacture a vaccine, but did not immediately export it because it wanted vaccines for its citizens first, says Amesh Adalja of the Johns Hopkins Center for Health Security in Baltimore, Maryland. "Most countries have laws enacted that allow the government to force manufacturers to sell domestically, and I don't see this changing," he says.

CEPI says that there is no agreement yet on the principles or rules for a fair allocation system incorporated into contracts that can be consistently applied and enforced. There is also no global entity responsible for ordering the manufacturing of vaccines on a global scale and paying for it.

"This is a challenge that must be urgently and collectively addressed by governments, global health leaders and regulators while COVID-19 vaccine development is continuing," says Mario Christodoulou, a communications manager at CEPI.

The WHO has tried to step in before to make sure that vaccine stockpiles are shared equitably, says Alexandra Phelan, of Georgetown University's Center for Global Health Science and Security in Washington DC. After the outbreak of H5N1 in countries such as China, Egypt and Indonesia, WHO member states adopted a resolution known as the Pandemic Influenza Preparedness (PIP) Framework. Under PIP, countries provide virus samples to a network of labs coordinated by the WHO, with the understanding that the organization would consider them in an as-needed basis to access a WHO stockpile of vaccines, diagnostics and drugs in the case of an influenza pandemic. But because PIP is designed for influenza, it doesn't apply to the current coronavirus outbreak.

Countries could agree on a framework similar to PIP for the current pandemic, but it is highly unlikely that a draft agreement would be ready in time for a World Health Assembly slated for May, at which member states would have to vote for it. And because there is so much SARS-CoV-2 already circulating, it's unclear whether this sort of agreement would work, because vaccine manufacturers can access virus samples from private labs, Phelan says.

It is possible that by the time a vaccine arrives, much of the world will already have been infected with the new coronavirus. Even in that case, however, many might want shots to boost immunity. And thinking ahead to ensure there's enough manufacturing capacity for vaccines in any future epidemic is still vital, Yaqub says.

"The concern for how to manufacture vaccines efficiently, reliably and safely is always going to be there," he says, "even if we can't get a coronavirus vaccine or we've managed to figure out other ways to deal with coronavirus."

Roxanne Khamsi is a science journalist based in Montreal, Canada.

Books & arts

MARIO TAMA/GETTY

The need for food banks has risen during the coronavirus pandemic.

Post-pandemic economic overhaul will take more than tweaks

As COVID-19 exacerbates inequalities, Thomas Piketty's analysis reads as timely, but inadequate. **By Ingrid Harvold Kvangraven**

The COVID-19 pandemic is exposing and exacerbating inequalities around the world. Read against this backdrop, economist Thomas Piketty's latest book is timely, but partial.

In *Capital and Ideology* (first published as *Capital et ideologie* in 2019), Piketty documents the global rise of inequality and critiques ideas that legitimize it. He builds on his bestselling 2013 book *Le Capital au XXIe siècle* (*Capital in the Twenty-First Century*), which spurred a public debate on growing gaps between the haves and have-nots in Europe and the United States. His latest work is important, especially because – before the pandemic – the London-based magazine *The Economist* had raised doubts about the extent to which inequality has really been rising. But in downplaying the roles of material interests, structures of production and capitalist dynamics, Piketty's analysis is concerning.

His argument is that societies always try to justify their imbalances, and that the prevailing justification rests on shaky foundations. He argues that differences in wages today are often justified by a "meritocratic fairy tale", in which people believe that the entrepreneurial earn wealth and those living in poverty simply need to work harder. But, of course, Western societies are not meritocratic. As Piketty

demonstrates, discrimination is common — based on status, race, gender and religion. In the COVID-19 pandemic, could our obvious dependence on undervalued work in sectors such as nursing, care of children and older people, grocery provision and delivery shift perception of the extent to which these workers deserve the low wages of their jobs, which are often precarious? I hope so.

Piketty discusses what he sees as the success of the period of social democracy in Europe and the United States in the 1950s to the 1970s, when the gap between the richest and poorest was narrower. He notes that most people who voted for social-democratic parties between 1950 and 1980 were workers, but that the vote has since shifted to the educated and middle class. Uneducated workers have thereby largely been left behind, paving the way for phenomena such as the election of US President Donald Trump and the United Kingdom's referendum on leaving the European Union.

However, as economist Michael Roberts has pointed out, the social-democratic period rested on compromises between capitalists, organized labour and the state, not on a coherent set of beliefs. What's more, Roberts says, the collapse of these alliances might have had more to do with plummeting profitability in the 1970s, which made it harder for social-democratic politicians to support workers.

War of words

Strikingly, Piketty does not recognize the political battle over ideas in academia, although this could help him to explain shifts since the 1970s, including economics departments squeezing out Keynesian and Marxist perspectives. Instead, Piketty simply draws a sharp line between knowledge production and politics. He labels his own empirical work "rational" and "unbiased", but his policy recommendations "ideological".

This is problematic. Economists' perceptions of their own analyses as being free of ideology often hinder open and democratic debate. The behavioural-economics work suggesting that the United Kingdom should not enter lockdown, which might have guided the UK government at the beginning of its COVID-19 response, is just one example. In that case, a particular way of seeing the economy — as composed of separate individuals responding rationally to incentives — was presented as an objective foundation for evidence-based policy that legitimized delays in social distancing. Yet such evidence cannot be considered purely objective, and in this case it contradicted World Health Organization recommendations.

MANOJ DHAKA/HINDUSTAN TIMES VIA GETTY

Farmers harvest wheat in India despite a nationwide lockdown.

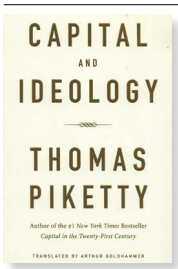
Piketty makes sweeping statements: he sees ideologies as social constructs with lives of their own, independent of what stakeholders stand to gain or lose. For example, he argues that one of the stated justifications for colonialism was the colonizers' idea of having a "civilizing mission"; this is true, but the prevailing motivation was without doubt the vast wealth to be acquired. Clarity here is essential for understanding the generation of massive global injustice. Similarly, Piketty does not provide convincing evidence that, as he claims, inequality in post-colonial countries such as South Africa is driven by ideas legitimizing chasms in opportunity, rather than, for example, the stubborn persistence of racist institutions.

It is ironic that Piketty nods frequently to Karl Marx while simultaneously ignoring key Marxist insights about dynamics such as the profit motive, unequal access to and ability to develop technology, and labour-squeezing cost-cutting. At times, it seems that Piketty simply equates capital with wealth, because he focuses both his analysis and his policy recommendations largely on wealth transfers. For example, rather than interrogating how we as society work, produce and consume, his solutions are biased towards redistribution without changing the core of the system.

This limits his capacity to explain global phenomena. This is clear in his view on the effects of trade liberalization: rather than exploring how the removal of barriers to imports in the 1980s led to a collapse of industry in the global south, Piketty focuses on the loss of income from tariffs. In the same vein, his proposals shy away from discussing the massive rebalancing of global finance and production that is necessary; instead, he focuses on aid transfers to governments, and taxation.

His policy proposals don't challenge our reliance on capitalist growth. Rather, they involve adjustments to the existing order, such as redistribution and the inclusion of employees on company boards. Therefore, the worry articulated in the United Kingdom's most right-leaning quality newspaper, *The Daily Telegraph*, that Piketty is back "more dangerous than ever", because of his vilification of entrepreneurs and billionaires, is in my view unfounded.

Despite its shortcomings, this book does have the potential to start an important debate about how to restructure society in a more egalitarian and ecologically sustainable way. If we are to exit the global depression brought about by the current pandemic with a system set for net-zero emissions, this will be more important than ever. But these debates must also involve more careful analysis of capitalist dynamics and the social relations of production.



Capital and Ideology
Thomas Piketty
Harvard Univ. Press (2020)

Ingrid Harvold Kvangraven is a lecturer in international development at the University of York, UK.
e-mail: ingrid.kvangraven@york.ac.uk

Elixirs for times of plague and bullion shortage

Spectacular alchemical scrolls record ideas of flux in times of massive upheaval — medical, social, economic and political. **By Jennifer Rampling**

Painted on a seventeenth-century parchment roll, a human-headed eagle devours its own wings. This bizarre figure is not the product of myth or divine revelation, but of chemistry — visual shorthand for the solidification of a once-volatile substance as it loses its ability to 'fly'.

The manuscript in which it appears, one of the Ripley Scrolls, is the centrepiece of an exhibition on alchemical imagery. Originally scheduled to open this month at Princeton University Library in New Jersey, it is now postponed until April 2021 in light of the current pandemic. As curator of the exhibition, and author of a related book out this year, I am vividly reminded of the economic and medical upheavals that alchemy was once intended to combat.

Fifteenth-century Europe ran low on bullion, partly owing to balance-of-payments deficits with Asia and the Middle East, sharpened by the local effects of war, famine and plague. Alchemists sought to resolve the shortage by transmuting base metals into gold and silver. To counter plague, they pursued medicinal elixirs. Their methods, disguised behind allegorical language and obscure, fantastical imagery, seem disconnected from those of modern science and medicine. Yet their images reveal a deep concern with understanding and capturing the dynamic nature of matter — a goal that still resonates and fascinates.

Philosophers' stone

The Ripley Scrolls rank among the most spectacular products of the alchemists' tradition. The oldest surviving example was produced in England in the late fifteenth century, although its exact provenance is unknown. The design was repeatedly copied, and now survives in one codex, or book, and 22 parchment and paper rolls, ranging from one to 7 metres long. Each is embellished with vibrant figures and verses, which set out the process for making the philosophers' stone and the elixir of life.

Ripley Scrolls were collected and studied by some of the most influential figures in English alchemy, including the Elizabethan astrologer and mathematician John Dee and the natural philosopher Isaac Newton, who made detailed notes on their design. In the seventeenth century, Elias Ashmole, a founding member of the Royal Society, owned five of the rolls and attributed the original to George Ripley, a canon from Yorkshire, who was famous for his 1471 poem *The Compound of Alchemy*.

Ripley might not actually have designed the original Scroll himself, but the images still shed light on how he and his contemporaries viewed matter and chemical change. The design opens with the figure of a philosopher holding a large glass vessel. Inside it, the first steps of his work are depicted in miniature allegorical scenes — incomprehensible to uninformed viewers, but packed with meaning for those in the know. The first shows a man and woman being attacked by humans and animals including a lion and a toad, signifying the 'death' (or dissolution) of metallic bodies.

> **"Fixed, solid matter might be signified by an earthbound toad, or a volatile substance by a bird or feathers."**

The complexities of alchemical imagery echo a deep problem for medieval natural philosophers: how to visualize the inner structures and workings of matter. Unlike models of the cosmos or natural-history specimens, medieval theories of matter did not lend themselves easily to either diagrammatical or naturalistic representation. Substances change state during chemical operations, losing one form while mysteriously acquiring another. Such dynamic processes, involving change at a level that cannot be grasped by the senses, are almost impossible to describe in a way that accurately conveys either matter's profundity or the minute incremental shifts through which it alters. The problem persists today, as evinced by modern attempts to represent the interior of the atom, or to model the structures of complex molecules.

For alchemical image-makers, one solution was to invoke analogies with other dynamic aspects of creation, from human reproduction to the motion of heavenly bodies. Byzantine Greek texts described ingredients and processes using metaphorical language (mercury could be 'dew' or 'seed of the dragon') that later offered rich resources for illustration. As medieval Muslim and Christian alchemists adopted this language, their imagery grew more sophisticated, developing major tropes such as the 'chemical wedding' of Sun and Moon — variously interpreted as sulfur and mercury or gold and silver.

Medieval writers did not assume that matter was made of particles. Europeans adopted an idea from Islamic alchemy, which saw metals in terms of two primordial constituents: hot, dry 'sulfur' and cool, running 'mercury' (principles that do not always correspond to the elements that share those names). At the same time, both Islamic and Christian philosophers followed Aristotle in defining substances according to bundles of properties: hardness, yellowness and heaviness, for instance, in the case of gold. Aristotle suggested that these attributes were carried by an underlying 'prime matter' that had no form of its own and existed in a potential, indiscernible state.

Allegorical dragons

The representation of properties, rather than structures, became a major feature of European alchemical imagery. In the fifteenth century, alchemists developed elaborate sequences of allegorical illustrations that allowed informed readers to decipher basic processes. Fixed, solid matter might be signified by an earthbound toad, or a volatile substance by a bird or feathers. A solvent could be a voracious lion or wolf.

Above all, alchemists strove to capture the fluid, penetrative properties of 'mercury'. The term described not just metallic quicksilver, but other products, including philosophical mercury — a solvent made from quicksilver. In several fifteenth-century illustrations, mercury is shown as a hybrid creature, half-human and half-serpent: a visual reference to its dual state as 'body' (used to denote metals) and 'spirit' (volatile substances). In the Ripley Scrolls, philosophical mercury is

Left: Princeton's Ripley Scroll. Above: A toad represents 'prime matter'.

Above left: A child of Sun and Moon. Above right: A bird eats its wings.

Above: The serpent of philosophical mercury dissolves gold and silver.

LEONARD SMETHLEY MS. ROLL/PRINCETON UNIVERSITY LIBRARY

variously depicted as a dragon eating a red toad (perhaps representing the dissolution of red lead) and as a snake-tailed woman in a tree. The latter is set in an alchemical rendering of the Garden of Eden, in which the mercurial serpent extracts the souls of Adam and Eve (gold and silver). Many of these images can be tied to specific processes, such as the production of a medicinal elixir based on lead compounds and distilled vinegar, popularized by Ripley.

Royal patronage

The Scroll was probably created as a gift to solicit patronage. The final section shows a plainly dressed figure – the author of the original – clasping a pilgrim's staff topped with a pen nib. In some versions, he faces a king, possibly the intended recipient. Henry IV made alchemy illegal in England in 1403–04 (reflecting concern over counterfeit coinage) but English monarchs could still permit individual alchemists to practise. My forthcoming book, *The Experimental Fire*, traces how rulers including Henry VI, Edward IV, Henry VIII and Elizabeth I issued licences or sponsored alchemical projects – including the production of multipurpose elixirs for transmutation and healing.

Although the basics of alchemical imagery remained consistent over centuries, the underlying chemistry inevitably changed. One consequence was a loss of meaning over time, as figures acquired generic identities. By 1720, proponents of the new discipline of chemistry had adopted their own forms of nomenclature and representation.

Alchemical images are now completely uncoupled from mainstream scientific practice. Nevertheless, they capture the colour and dynamism of chemical transformations in ways seldom achieved by their more respectable textbook cousins.

The Ripley Scrolls continue to attract attention when displayed, including at the British Library's wildly successful exhibition Harry Potter: A History of Magic, in 2017–18. In December 2017, the last Scroll in private hands was purchased in London by Princeton University for more than £500,000 (US$620,000). I hope that it will finally go on display next April, alongside an earlier copy, at Princeton's exhibition Through a Glass Darkly: Alchemy and the Ripley Scrolls. Their chemistry might have transmuted, but the legacy of the alchemical philosophers is still being unrolled.

Jennifer Rampling is assistant professor of history at Princeton University in New Jersey. She is author of *The Experimental Fire: Inventing English Alchemy, 1300–1700* and curator of Princeton's forthcoming exhibition on the Ripley Scrolls.
e-mail: rampling@princeton.edu

E. Margaret Burbidge
(1919–2020)

Astronomer and co-discoverer of evidence that elements are made in stars.

EMILIO SEGRE VISUAL ARCHIVES/AMERICAN INSTITUTE OF PHYSICS/SPL

Margaret Burbidge was the director of the Royal Greenwich Observatory from 1972 to 1973.

In 1957, a landmark paper brought together theoretical and observational studies supporting the idea that all the heavier elements in the Universe, from carbon to uranium, were synthesized in stars through nuclear fusion. The first author alphabetically was the observational astronomer Eleanor Margaret Burbidge (née Peachey), always known as Margaret, who has died aged 100.

Her co-authors were her husband Geoffrey Burbidge, the nuclear physicist William Fowler and the astronomer Fred Hoyle, proponent of the steady-state model of the Universe. The paper was so influential that generations of astrophysicists called it B^2FH for short, quipping that the early Universe made hydrogen and helium, but Burbidge, Burbidge, Fowler and Hoyle made all the rest.

Burbidge, often working with Geoffrey, also measured the rotation and masses of spiral galaxies, documented the very large redshifts of quasars and contributed to the development of the Faint Object Spectrograph, an instrument launched aboard the Hubble Space Telescope in 1990.

One of the Burbidges' collaborators on the galaxy-rotation studies, Vera Rubin, later detected motion in the far outskirts of spiral galaxies, and so discovered evidence for dark matter. This was only one of the many ways in which Margaret Burbidge blazed a trail for female astronomers. Having once been refused access to observatories because of her gender, she later became the first female president of the International Astronomical Union's commission on galaxies, director of the Royal Greenwich Observatory in London and president of the American Astronomical Society.

Burbidge was born in Davenport, near Manchester, UK, the daughter of a chemistry lecturer and his former student. Her parents encouraged her interest in science with gifts of a telescope, suitable books and a chemistry set. After attending all-girl schools, she entered University College London, where she completed a PhD in 1943 on the spectrum of the star Gamma Cassiopeiae. She enjoyed greater scientific independence and responsibility at the University of London Observatory than she might have otherwise, because many of the male academic staff were engaged in war work.

Geoffrey turned up as a fellow student in 1947, and they married the following year. They bounced between several US and UK institutions, looking for somewhere that would give them both jobs and access to telescopes. At this time, they analysed the spectra of many stars with unusual surface compositions. These sometimes came from upward mixing of the products of nuclear reactions and so became some of the raw material for the B^2FH analysis.

They met Hoyle and Fowler while in Cambridge, UK, for one year in the mid-1950s. Hoyle had started thinking about the products of nuclear reactions in stars before the Second World War. Fowler's nuclear-physics group at the California Institute of Technology (Caltech) in Pasadena had already measured cross-sections for many of the reactions needed to build heavy elements from hydrogen and helium; Geoffrey Burbidge wrote most of the 100-page 1957 paper. Astrophysicist Alastair G. W. Cameron simultaneously identified nearly the same processes, but tends to get left out of the story because his paper was initially classified.

During the Burbidges' first stay in southern California, Geoffrey the theorist had an appointment at Mount Wilson Observatory, entitling him to apply for observing time, which women could not. Margaret the observer had a theory position at Caltech. He applied, and she did the observing. Later, after they settled at the University of California, San Diego (UCSD), she had unquestioned access to the Lick Observatory and its 3-metre telescope. After the 1963 discovery of quasars, she used much of her time to study their spectra, contributing to the discovery of absorption lines indicating the presence of intergalactic gas clouds, and measuring the record redshifts of some of the most distant objects then known.

To many, her observations provided evidence of the Big Bang, now almost unanimously accepted as the origin of the Universe. However, both she and Geoffrey followed Hoyle in supporting the steady-state theory. From the start, Geoffrey, and later Margaret, felt that their observations showed that quasars were much closer than their redshifts indicated, and that the data could not be used to rule out steady-state cosmology.

The Burbidges' return to the United Kingdom in 1972–73 was brief. All previous directors of the Royal Greenwich Observatory had also held the title of astronomer royal. A change of policy at the time of her appointment made this a purely honorary role, separate from the directorship. Was it a gender issue? Perhaps; perhaps not. Nevertheless, she resigned after 18 months and they returned to UCSD, where she remained for the rest of her long working life. She took US citizenship following her election as American Astronomical Society president for 1976–78.

In 1972, she declined the society's Annie J. Cannon prize, awarded only to women, writing that it was no longer needed or appropriate. Subsequently, the society restricted the prize to early-career women who chose to apply for it. As president, she introduced Cecilia Payne-Gaposchkin as the first woman to receive the society's Russell Lecturer award for a lifetime of excellence in astronomical research. Burbidge became the second in 1984.

If there was anyone who got to know Margaret and didn't like her for her intelligence, charm and patience, I never met them.

Virginia Trimble is professor of physics and astronomy at the University of California, Irvine.
e-mail: vtrimble@uci.edu

Obituary

Jennifer Clack

(1947–2020)

Palaeontologist who described how vertebrates moved from water to land.

Jennifer A. Clack made groundbreaking fossil discoveries of the emergence of animals with backbones out of the water onto the land. Her work lifted the study of the transition from fishes to tetrapods from palaeontological obscurity to star status, ranking alongside such favourites as the origin of birds or the evolution of the human lineage. A happy convergence of brilliance, tenacity, opportunity, generosity and modesty enabled Clack (née Agnew) to rejuvenate an entire research field. She died on 26 March, aged 72.

Clack amassed an unprecedented trove of several hundred tetrapod fossils from the Devonian (419 million to 359 million years ago) and Carboniferous periods (359 million to 299 million years ago). She did so through a combination of fieldwork in southern Scotland and two expeditions to Greenland, the latter following up a chance discovery in a museum drawer. These strange animals resembled large salamanders or small crocodiles but they retained fish-like characteristics such as tail fins. Her papers describing and interpreting them shed light on the evolution of the body plan of land vertebrates. She broke open the entrenched positions of the previous generation of researchers.

Clack was born in 1947 in Manchester, UK. She studied zoology at the University of Newcastle upon Tyne, graduating in 1970. One of her lecturers, Alec Panchen, was re-evaluating Carboniferous tetrapod fossils discovered during the nineteenth century. Clack was intrigued, but Panchen was unable to offer her a PhD position. Instead, Clack did a one-year graduate certificate in museum studies at the University of Leicester, which led to a job at the City of Birmingham Museum and Art Gallery. During this time, she met fellow biker and fossil enthusiast Rob Clack. They married in 1980.

In 1978, Clack was finally able to begin a PhD in Panchen's lab, working on a specimen of the large, salamander-like Carboniferous tetrapod *Pholiderpeton*. It proved to have an ear bone (stapes) shaped like a butterfly and probably no eardrum, defying expectations that these early tetrapods would have had ears like those of frogs, with an eardrum and a rod-like stapes. Clack proposed that the 'middle ear' of the earliest tetrapods was much more primitive, still a small gill opening as in fishes, and that the drum evolved later, a hypothesis that is now widely accepted.

MICHAEL COCKERHAM

In 1981, Clack secured a permanent position at the University Museum of Zoology in Cambridge, UK, where she remained. In 1986, she made a chance discovery that defined her career: in the Sedgwick Museum in Cambridge she found fossils of Late Devonian tetrapods from eastern Greenland, collected in the 1960s.

Only three Devonian tetrapods were then known, two of them (*Ichthyostega* and *Acanthostega*) also from eastern Greenland.

> **"Her work lifted the study of the fish–tetrapod transition from palaeontological obscurity to star status."**

The *Ichthyostega* material was quite extensive; *Acanthostega* was little more than a name attached to half a skull. In the Sedgwick, Clack found several skulls of *Acanthostega*, with sections of vertebral column attached, the tight fit of the pieces showing that the tetrapods had been lying packed together in the rock.

In June 1987, Clack set off for Greenland with Rob, myself (her first PhD student) and two colleagues from Copenhagen to find more specimens. The results exceeded her wildest expectations. She returned with the largest haul of tetrapod fossils ever recovered in one season from the Devonian deposits in Greenland. Back in Cambridge, Clack recruited two new team members – a postdoc, Michael Coates, and a fossil preparator, Sarah Finney – and set to work.

A series of groundbreaking papers poured forth describing *Acanthostega* specimens, many of which showed new and unexpected features that were to overturn established views of the transition from water to land.

The paper with the greatest impact – popular and scientific – showed that the feet of *Ichthyostega* and *Acanthostega* had, respectively, seven and eight toes apiece, rather than the canonical five (M. I. Coates and J. A. Clack *Nature* **347**, 66–69; 1990). The paper received the ultimate pop-science accolade: the US palaeontologist Stephen Jay Gould devoted one of his regular essays in *Natural History* magazine to it. A follow-up expedition to Greenland in 1998 found more material, especially of *Ichthyostega*.

During the last decade of her life, Clack increasingly turned her attention back to the Carboniferous tetrapods. She worked especially on 'Romer's Gap' – the 30-million-year break in the fossil record between the Late Devonian forms and the more advanced tetrapods of the mid-Carboniferous. She suspected that this lacuna might be a sampling artefact. She assembled a team of researchers to investigate sediments in northern England and southern Scotland, called the Ballagan Formation, which represent the earliest Carboniferous (around 359 million to 347 million years ago).

These sandstones and mudstones promptly started yielding fossils of previously unknown tetrapods. Six genera have been described so far and more material awaits description. They look set to revolutionize our understanding of the early diversification of tetrapods by filling in the wide morphological gap between very primitive ones, such as *Acanthostega*, and the more modern tetrapods of the Carboniferous.

Amid all this triumph, and having seen her book *Gaining Ground* (2002, 2012) become the standard text on the origin of tetrapods, Clack noticed the first symptoms of cancer. She continued working until her last few days. Clack leaves behind a vibrant research area, characterized by collegial openness that she did much to foster. Her death leaves a very big void.

Per Ahlberg is professor of evolutionary organismal biology at the University of Uppsala, Sweden.
e-mail: per.ahlberg@ebc.uu.se

Comment

EVA PLEVIER/REUTERS

A farmer in the Netherlands with tonnes of potatoes meant for food outlets that have closed because of the coronavirus pandemic.

Without food, there can be no exit from the pandemic

Máximo Torero

Countries must join forces to avert a global food crisis from COVID-19.

The coronavirus pandemic has laid many things bare, none more so than how interconnected our world is. The impact of globalization is most obvious in the stuttering supply chains that threaten food security worldwide. Maintaining or reweaving these webs is going to take technology, innovation and political determination.

As chief economist at the Food and Agriculture Organization of the United Nations (FAO), I fear that few countries have recognized that their measures to contain the virus and buffer economic shocks must be adjusted to keep food flowing. Without food, there can be no health. The policy prescriptions are straightforward, and isolationism can form no part of it. Countries must work together, not throw up trade walls and bar essential workers from crossing borders.

Global food-supply chains are already buckling. In India, farmers are feeding strawberries to cows because they cannot transport the fruit to markets in cities. In Peru, producers are dumping tonnes of white cocoa into landfill because the restaurants and hotels that would normally buy it are closed. And in the United States and Canada, farmers have had to pour milk away for the same reason. Legions of migrant workers from Eastern Europe and North Africa are trapped at borders, instead of harvesting on the farms of France, Germany and Italy. The United States, Canada and Australia all rely heavily on seasonal farmworkers who are unable to travel because of virus restrictions, including the suspension of routine visa services by some embassies. There are also concerns that foreign workers could import cases of infection. Crops are rotting in the fields.

Fortunately, cereal harvests are expected to be good this year. Already, the world's stockpile of maize (corn) is more than twice what it was in 2007 and 2008, when severe droughts created food shortages in key exporting countries, leading to a global food crisis. Rice and soya-bean stockpiles have also increased over this period, by around 80% and 40%, respectively.

But the bounty will not help to avert food shortages if countries cannot move food from where it is produced to where it is most needed. Ships laden with cereals, fresh fruit and vegetables are docking late and their crews cannot disembark. So perishables, unable to reach wholesale markets in time, are going to waste. Wheat prices have jumped by 8% and rice prices by 25% compared with those of March last year. Meanwhile, panic buying across the world is creating more waste and affecting the quality of diets as people struggle to access fresh food. Global action on food was a challenge even before COVID-19. That countries and regions are experiencing the pandemic at different times and in different ways — from China, to Europe, the United States, India and now Africa — has created an ethos of nations acting only for themselves.

Chain reactions

That has led to chaotic chain reactions. Earlier this month, Russia, the world's leading wheat exporter, limited wheat exports for three months to ensure that local supplies were sufficient. Although the disruption is expected to be minimal (see, for example, ref. 1), the gesture set alarm bells ringing elsewhere. It was a decision driven by a confluence of events, including the sharp drop in oil prices — this weakened the rouble against the dollar, which in turn bumped up local prices of wheat. It is the same course that Vietnam took with paddy rice in March, which is why rice prices spiked.

The pandemic has emboldened divisive arguments — such as that open borders have enabled the virus to spread, that refugees and immigrants must be kept out, and that outsourcing should end. But such political positions ignore how much nations depend on each other for staple ingredients, pesticides, fertilizers, animal feed, personnel and expertise.

What happens next depends on whether nations resist isolationist pressures. I urge them to commit to not imposing export restrictions in response to the pandemic. Instead, they should agree to eliminate tariffs and taxes to compensate for local price increases caused by currency devaluation. And they should designate workers at ports and on farms as essential personnel, protect the health of these people and ensure that they can travel and continue to work.

Collaboration is possible. The agriculture ministers of 25 Latin American and Caribbean countries signed an agreement this month to work together to guarantee food supplies in the region. Such a political declaration can pave the way for real progress. And governments and investors can benefit from more transparency and information than ever before on market conditions, through tools such as the Agricultural Market Information System (www.amis-outlook.org), which can reduce uncertainty.

Smoothing the shocks

At the FAO, we are focusing on mitigating the virus's impact on the activities that deliver produce to people, using evidence and lessons learnt from past crises. This includes information about food-price increases and volatility[2], and how access to food and nutrition was affected during recent outbreaks such as that of Ebola[3]. Using big data, we monitor trade and collect information on logistical issues, assess how problems have been resolved and then signal the outcome to the market to reduce uncertainty (see https://datalab.review.fao.org). For example, we know that the main delay in shipping happens during cargo unloading, which now takes three days instead of one because of labour restrictions at ports. The delay is expensive for exporters, but they make up for it with the gains from exchange rates. So global shipping is working.

"Nations depend on each other for staple ingredients, pesticides, fertilizers, animal feed, personnel and expertise."

We also track news in multiple languages to see how the pandemic is affecting food and agriculture. This helps countries to make policy decisions. We work with developing nations to boost food supply by analysing their agro-ecological conditions and advising when and where to plant and harvest their key commodities. We forecast how various aspects of the agricultural sector could be affected by COVID-19 — from labour and decreased demand because of falling incomes to exchange rates and inflation[4].

What the pandemic has underscored is that the world must use its land and water resources sustainably, to grow essential, nutritious food in a more resilient way. One way of doing this is to cut food loss[5]. The world squanders about US$400 billion of food annually — an amount that could feed around 1.26 billion people a year. The wastage is equivalent to 1.5 gigatonnes of carbon dioxide emissions. (Compare that to the roughly 33 gigatonnes emitted in 2019 to produce the world's energy.) Another priority is better treatment for smallholders and migrant workers, who form the backbone of farming. For example, small-scale operations need access to markets and help to increase productivity and incomes, which goes far beyond mere subsidies.

The pandemic is an opportunity to hit the reset button, with scientists and social scientists playing an important part. And innovation is happening: China is investing in drones, unpiloted vehicles and other agriculture technologies to reduce human contact. In Africa, mobile phones are improving access to markets, prices and weather data, as well as facilitating money transfers[6]. Peru is seeing the benefits of innovative legislation that has formalized the farm labour force and directly linked it to the seasonality of crops. The government now knows which farmers are affected by the lockdown and can ensure they receive the necessary support. Let us seize these huge opportunities collectively.

It is precisely because the coronavirus doesn't respect borders that global cooperation is the only shot at defeating it. The people who are working on vaccine trials, health care, drug discovery and economic recovery must all still eat. We can either stand together or many millions will starve separately.

The author

Máximo Torero is the chief economist of the Food and Agriculture Organization of the United Nations in Rome, Italy.
e-mail: maximo.torerocullen@fao.org

1. Bouët, A. & Laborde Debucquet, D. *Economics of Export Taxation in a Context of Food Crisis. IFPRI Discussion Paper 00994* (International Food Policy Research Institute, 2010).
2. Torero, M. In *Food Price Volatility and Its Implications for Food Security and Policy* (eds Kalkuhl, M., von Braun, J. & Torero, M.) 457–510 (Springer, 2016).
3. World Bank. *The Economic Impact of the 2014 Ebola Epidemic* (World Bank, 2014).
4. Schmidhuber, J., Pound, J. & Qiao, B. *COVID-19: Channels of Transmission to Food and Agriculture* (FAO, 2020).
5. Food and Agriculture Organization. *The State of Food and Agriculture 2019: Moving Forward on Food Loss and Waste Reduction* (FAO, 2019).
6. Trendov, N. M., Varas, S. & Zeng, M. *Digital Technologies in Agriculture and Rural Areas — Status Report* (FAO, 2019).

Correspondence

COVID-19: sample for future analysis

Resources for COVID-19 testing in many parts of the world are still limited. We suggest that future value could be realized if samples were to be widely taken now and saved for analysis as more resources become available.

Besides diagnostic screening, this sample analysis alone — or in combination with mobile and other data — could provide insight into incubation periods, the relationship between transmissibility and symptoms, disease progression in individuals, and the effectiveness of different mitigation measures.

Simple buffers that are commercially available or readily prepared in the laboratory (for details, see S. Menke *et al. Front. Microbiol.* https://doi.org/dr2s; 2017) can indefinitely preserve nucleic acids in samples of saliva, nasal mucus and other biological material at room temperature. This form of sample preservation would obviate the need for refrigeration under difficult field or hospital conditions (see M. Camacho-Sanchez *et al. Mol. Ecol. Resour.* **13**, 663–673; 2013).

David S. Thaler Biozentrum, University of Basel, Switzerland. david.thaler@unibas.ch

Marc Lipsitch Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, Massachusetts, USA.

Lockdown: keeping research on track

As researchers in one of the regions most severely affected by COVID-19, we have been able to continue our work by using a rigid system of 24-hour shift-working that complies with the Italian government's strict lockdown conditions.

Under this scheme, group leaders allocated volunteers from their teams to one of three shifts, taking into account ethical issues around sample handling and the cost and relevance of ongoing experiments. The shifts run from 05:00 until 12:00, 12:30 to 17:30 and 18:00 to 04:00; time between shifts is used to sanitize exposed surfaces. Laboratory occupancy at any one time is 10–15% of normal, optimizing safety. For experiments that take longer than a single shift, researchers modify their protocols — for example, by freezing samples — or ask colleagues on the next shift to take over.

People must self-report their health status before coming in: anyone knowingly exposed to or showing symptoms of coronavirus infection is excluded. And those sharing a work space with an individual who develops symptoms, or has come into contact with a symptomatic person, are immediately quarantined and medically tested if necessary (the department is a certified centre for SARS-CoV-2 diagnostic testing).

Gioacchino Natoli, Saverio Minucci, Pier Giuseppe Pelicci European Institute of Oncology, Milan, Italy. gioacchino.natoli@ieo.it

Revamp economics for global fixes

Hetan Shah rightly argues that social insight is needed to help meet the challenges of the next decade (*Nature* **577**, 295; 2020). The important question is whether social science — and economics, in particular, as the queen of the social sciences — is up to the task of informing government policy.

After spending two decades drilling down on economic theory, I can testify that the field lags badly behind the development of hard science and engineering. Our universities must bring into the present the essential considerations of economics that were formed in the mid- to late-nineteenth century — when economic behaviour was explained in terms of the 'marginal revolution' (W. Jaffé *Hist. Polit. Econ.* **4**, 379–405; 1972). To do so, they need to incorporate the important advances made since then in basic temporal and expectational economics.

Mathematical economics is not a sufficient condition for meeting the next decade's challenges — but it is certainly a necessary one.

Thomas Chamberlain Los Angeles, California, USA. thomas.e.chamberlain@gmail.com

Scientists think on the move

Your 'Where I Work' series (see, for example, *Nature* **580**, 158; 2020) reminds me that travel itself can be scientifically productive. Travel can free the mind from its routine tasks, opening it up to new ideas. The polymerase chain reaction, for example, came to the late Kary Mullis while driving late one night from Berkeley to Mendocino in California. A lynchpin in today's screening of thousands for COVID-19, it earned him the Nobel Prize in Chemistry in 1993.

Other Nobel-worthy ideas seeded on the move include Subrahmanyan Chandrasekhar's insight into the maximum mass of a white dwarf star, and John Robert Schrieffer's mathematical description of the ground state of superconducting atoms. Chandrasekhar was on a ship from India to England when he had his brainwave at the age of 19; it won him the Nobel Prize in Physics in 1983. Schrieffer's came to him on New York City's subway, earning him a Physics Nobel in 1972.

Imagined travel, too, has contributed to discoveries. A notable example is Albert Einstein's 1907 'elevator' thought experiment, which inspired his general theory of relativity. Another is mentioned in Dante Alighieri's *The Divine Comedy*, when the poet describes Galilean relativity in the context of flight on the back of a monster (L. Ricci *Nature* **434**, 717; 2005).

Eric L. Altschuler Metropolitan Hospital, New York, USA. altschue@nychhc.org

News & views

Sociology

Monitoring global education inequality

Monica Grant

Tools have been developed to project inequalities in education around the world to 2030. They reveal that overall inequality will decline, but that all world regions will fall short of achieving universal secondary education. **See p.636**

Increased years of schooling have been linked to better health and survival[1], slower population growth[2] and greater economic growth[3]. Because of its importance, access to "inclusive and equitable quality education" was included as one of the Sustainable Development Goals (SDGs) ratified by the United Nations General Assembly in 2015 (see go.nature.com/3ana8ob). The SDGs are an ambitious set of international development targets to be achieved by 2030. Friedman *et al.*[4] provide evidence on page 636 that, although most nations are projected to achieve near-universal primary education by 2030, large inter-regional disparities in the rates of secondary-school completion will persist.

The authors set out to assess whether countries are on track to achieve the SDGs for education by 2030. They assembled a database of 3,180 nationally representative censuses and surveys from 195 nations and territories. This database is an improvement on previous efforts to monitor education, which relied either on data back-projected from a single time point[5] or on a database derived from one-fifth as many data sources[6]. Friedman *et al.* developed a model that combines all the data sources in their data set and extrapolates single-year estimates of educational attainment for all populations, separately by sex and country, from 1970 to 2018. The model then uses this information to project future trends in educational attainment for individual countries or territories, and for seven 'major world regions' chosen by the authors, which include high-income countries, sub-Saharan Africa and Eastern Europe and central Asia.

Friedman and colleagues conclude that most countries are in line to achieve near-universal levels of primary-school attainment by 2030 — that is, for almost 90% of children to complete 6 years of education (Fig. 1). The exceptions to this trend are places such as Afghanistan, Papua New Guinea and parts of northern sub-Saharan Africa. By contrast, progress towards near-universal secondary attainment (12 years of schooling) is more uneven. Only 61% of young adults aged 25–29 years old are expected to have completed secondary school by 2030, and no major world region is expected to reach near-universal levels of secondary-school attainment. Furthermore, access to tertiary schooling is expanding faster in some world regions than in others; as a result, disparities are expected to increase until 2030.

The authors also find that gaps in educational attainment between men and women are expected to have changed substantially by 2030. In 1970, men achieved significantly more years of education than did women in 142 countries. By 2018, this had narrowed to 27 countries, and by 2030 it is projected to be just 4. Moreover, in 18 nations and regions, women are expected to achieve significantly higher mean years of schooling than men. This changing gap is largely attributable to girls' gains in primary schooling.

Next, Friedman *et al.* developed a metric for monitoring educational inequality within countries or territories — average interpersonal difference (AID), which measures the average difference in educational attainment between any two individuals in a population in a given year. The AID provides a different perspective from those of other commonly used metrics.

For example, consider the Gini coefficient, which is perhaps the most commonly used indicator of inequality. Under this coefficient, inequality is highest when education is concentrated in the hands of a few. As access to education expands, the Gini coefficient declines.

By contrast, the AID equates inequality with heterogeneity in educational attainment within a population. When education is concentrated in the hands of the few, the AID is

ZAKIR HOSSAIN CHOWDHURY/NURPHOTO/GETTY

Figure 1 | Primary schoolchildren in Dhaka, Bangladesh. Friedman *et al.*[4] analysed the number of years of schooling obtained by children in some 195 nations and territories between 1970 and 2018, and modelled predicted changes to 2030, to assess whether the world will meet the Sustainable Development Goals for education set by the United Nations.

low because most people have the same low level of educational attainment. As access to schooling expands, inequality as measured by the AID rises because the population now contains many people who have no education and many who have several years of schooling. As school enrolment becomes universal and members of the population begin to achieve similarly high levels of education, the AID declines again. From this metric, Friedman and colleagues conclude that global educational inequality peaked in 2017 and is projected to decline until 2030.

Even though this project involves an impressive volume of data, it is still limited by problems of data scarcity. Whereas some high-income countries, such as France and Germany, contribute more than 55 data points to the model, the time series for many resource-constrained and small-population countries or territories are extrapolated from fewer than 5 data points, or rely on data last collected in or before 2008. Although the validity of the model was evaluated by checking how well its predictions matched real data across many simulations in which one subset of data had been removed, there is no way of assessing how well it estimates attainment trajectories for places such as Malaysia, for which no data were available after 2003. The authors leverage regional trends to inform analyses of countries or territories for which data are scarce, but the results should be interpreted with caution.

The smoothed trajectories of predicted change in the study also hide the profound and often sudden impact of education policies on schooling. The authors note nonlinearities in the rates of change consistent with a sudden increase in schooling, which might result from the elimination of school fees or an increase in the years of compulsory schooling. The recent expansion of free secondary education in many lower-income countries has the potential to further advance progress towards the SDG education goals, beyond what is currently predicted by Friedman and colleagues' model.

Ultimately, we can monitor only what we can measure: we track trends in educational attainment and in gaps between the sexes because those are the data that exist. Socio-economic gaps in schooling are now substantially larger than are gender gaps in most world regions[7], but sufficient data on the socio-economic status of students are scarce. Likewise, almost three-quarters of countries have inadequate data with which to monitor progress in learning outcomes (such as mathematics or reading skills), rather than merely in years of schooling (see go.nature.com/39kd4o1). Global commitments to inclusive and equitable quality education run the risk of failing to achieve their true goals when we lack the data to properly track progress.

Monica Grant is in the Department of Sociology, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.
e-mail: grantm@ssc.wisc.edu

1. Marmot, M., Friel, S., Bell, R., Houweling, T. A. J. & Taylor, S. *Lancet* **372**, 1661–1669 (2008).
2. Abel, G. J., Barakat, B., Samir, K. C. & Lutz, W. *Proc. Natl Acad. Sci. USA* **113**, 14294–14299 (2016).
3. Lutz, W., Cuaresma, J. C. & Sanderson, W. *Science* **319**, 1047–1048 (2008).
4. Friedman, J. *et al. Nature* **580**, 636–639 (2020).
5. Goujon, A. *et al. J. Demogr. Econ.* **82**, 315–363 (2016).
6. Barro, R. J. & Lee, J. W. *J. Dev. Econ.* **104**, 184–198 (2013).
7. Jones, G. W. & Ramchand, D. S. *J. Int. Dev.* **28**, 953–973 (2016).

This article was published online on 15 April 2020.

Organic chemistry

Methyl groups make a late entrance

Emily B. Corcoran & Danielle M. Schultz

The addition of a methyl group to a drug molecule can greatly alter the drug's pharmacological properties. A catalyst has been developed that enables this 'magic methyl effect' to be rapidly explored for drug discovery. **See p.621**

Developing a small-molecule drug requires iterations of building and testing new compounds to find one that strikes the right balance of pharmacological properties. The process typically takes more than 10 years and costs billions of dollars, because, for every 5,000 compounds made and tested, only one will become an approved drug[1,2]. Indeed, a high-school basketball player is twice as likely to end up playing in the US professional league as any single compound tested in a drug-discovery programme is to become a marketed drug (see go.nature.com/2v8pnfm). One approach to accelerating drug discovery is late-stage functionalization, in which previously prepared test compounds are decorated with new atoms in the hope of favourably adjusting their pharmacological properties. On page 621, Feng *et al.*[3] report an outstanding advance towards this long-standing and historically challenging strategy.

> **"This work is a superb example of a symbiotic collaboration between academia and the pharmaceutical industry."**

Introducing just one cluster of atoms (a functional group) into a drug molecule can drastically alter the molecule's properties. For instance, adding a methyl group (CH_3, one of the smallest functional groups) can enhance a compound's binding affinity for its biological target more than 1,000-fold, a phenomenon termed[4] the 'magic methyl effect'. This is because the installation of a methyl group alters the shape of the molecule such that it can readily nestle inside a targeted protein's active site, akin to how an ergonomic computer mouse fits snugly in the palm of your hand.

However, making even small adjustments to molecules is frequently a major undertaking, one that effectively requires chemists to break apart the entire structure and reassemble a dozen or more smaller pieces for each change. Imagine how much time and money it would cost if adding a new window to your home required the entire house to be taken apart and rebuilt from scratch. Chemists working in drug discovery regularly have to do this with their molecules.

Late-stage functionalization has therefore emerged as a desirable approach to accelerate drug discovery[5,6]: much as a construction crew saws through existing walls to insert new windows, chemists aspire to cut through existing chemical bonds to insert new functional groups into molecules. C–H functionalization, a type of reaction that converts ubiquitous carbon–hydrogen (C–H) bonds in complex molecules into alternative functional groups, has garnered much attention for this purpose. Feng *et al.* report a substantial advance in this area with the design of a metal catalyst that cuts through specific C–H bonds to insert methyl groups, thus allowing the magic methyl effect to be explored in myriad complex and drug-like compounds.

Selective late-stage C–H functionalization is constantly used in nature. For example, iron-based metalloenzymes known as cytochrome P450s (CYP450s) are omnipresent throughout the animal kingdom because of their crucial role in regulating metabolism[7,8].

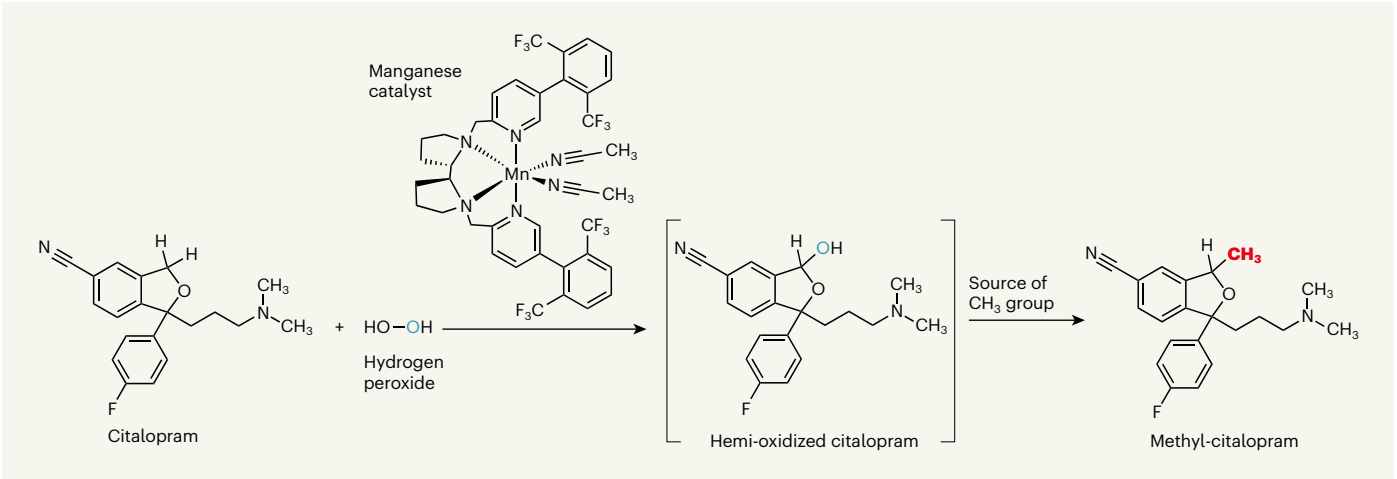


Figure 1 | Late-stage methylation of biologically relevant targets. Feng *et al.*[3] report that a highly tuned manganese (Mn) catalyst enables methyl (CH_3) groups to be incorporated at specific sites into complex molecules, particularly those that have structures typical of drugs. The manganese catalyst inserts an oxygen atom from hydrogen peroxide into the carbon–hydrogen bond that the human body can most easily metabolize, yielding a reactive hemi-oxidized intermediate (square brackets indicate that the intermediate is formed *in situ* and is not isolated) that is poised for reaction with a methyl-group source. The reaction was used successfully on 38 biologically active molecules, including the antidepressant citalopram. It could therefore be used to rapidly explore the magic methyl effect — a phenomenon in which the addition of a methyl group to a drug molecule greatly enhances the molecule's pharmacological properties.

The iron atom of a CYP450 binds to biologically active molecules and triggers their metabolism by inserting oxygen into C–H bonds to form double carbon–oxygen (C=O) bonds, a type of C–H functionalization. The elaborate enzyme architecture around the iron centre tames the metal's otherwise rampant catalytic activity, thus allowing these reactions to proceed precisely and specifically, such that only those substrates that fit in the enzyme's pocket are oxidized.

Feng and colleagues are part of a research group that has long been interested in making ligand molecules that mimic the CYP450-enzyme architecture, in the hope of broadening the ability of iron complexes to transform C–H bonds into C=O bonds in diverse substrates, using hydrogen peroxide as the source of oxygen[9]. Scientists from that group had previously made great strides in taming the reactivity of iron complexes for C–H functionalization, but even the best catalysts proved promiscuous (they reacted at many different C–H bonds, rather than at just one) and could not be used in the presence of many functional groups commonly found in drug-like molecules. The same research group had therefore also investigated manganese — iron's less-oxidizing neighbour in the periodic table — as an alternative metal centre for catalysts that oxidize specific C–H bonds in complex molecules[10].

Feng *et al.* hypothesized that a less-oxidizing manganese catalyst would target the C–H bonds that are most easily metabolized on drug-like molecules. Moreover, they thought that the oxidation reaction could be halted midway to produce a hemi-oxidized intermediate, into which a methyl group could be inserted (Fig. 1). This group would essentially block the molecule's metabolic degradation, invoking the magic methyl effect.

The challenge with this approach is that the hemi-oxidized intermediate is more readily oxidized than is the starting material — so, if the oxidation reaction were a train, it would be a non-stop service to a C=O bond. To circumvent this complication, Feng *et al.* tuned the reaction conditions to contain the precise amount of catalyst and hydrogen peroxide needed to deliver the hemi-oxidized intermediate, effectively pulling the train into a station en route to the C=O terminus. The resulting hemi-oxidized species can then be seamlessly transformed into a methyl group under a variety of conditions, depending on the functional groups present in the rest of the molecule.

Feng and colleagues' work is a superb example of a symbiotic collaboration between academia and the pharmaceutical industry, with cutting-edge chemistry being used to solve real-world problems. The industrial influence is evident throughout the work: the molecules selected to demonstrate this methodology accurately reflect the types frequently encountered in drug development. More specifically, the authors report that 38 biologically relevant targets (drugs, natural products, peptides and steroids) and their building blocks undergo the new reactions with excellent selectivity and functional-group tolerance.

For more than a century, drug discovery focused mainly on small molecules. However, the field is now turning to more-elaborate molecules, such as peptides, which can potentially target complex biological targets with high specificity. Peptides are usually stitched together from amino acids in a linear sequence of reactions. The functional groups that provide the structural diversity of peptides are built into the amino acids, and are therefore introduced at each step of the sequence. Some of the groups in a target peptide are inevitably installed in the first step, and can be changed only by running the whole sequence again, but using a different amino acid at the start.

Feng *et al.* upend this norm by demonstrating that methyl groups can be installed on a tetrapeptide (a peptide built from four amino acids) at the end of the synthetic sequence. Further extension of this chemistry to more-complex linear and macrocyclic (ring-forming) peptides would be game-changing for drug discovery. Continued breakthroughs on complex catalytic processes in the spirit of Feng and colleagues' work might finally enable medicinal chemistry to cruise at the same speed as biological research.

Emily B. Corcoran is in Process Research and Development, Merck & Co., Inc., Boston, Massachusetts 02115, USA.
Danielle M. Schultz is in Process Research and Development, Merck & Co., Inc., Kenilworth, New Jersey 07033, USA.
e-mails: emily.corcoran@merck.com; danielle.schultz@merck.com

1. Pritchard, J. F. *et al. Nature Rev. Drug Discov.* **2**, 542–553 (2003).
2. *Biopharmaceutical R&D: The Process Behind New Medicines* (Pharmaceut. Res. Manufact. Am., 2015).
3. Feng, K. *et al. Nature* **580**, 621–627 (2020).
4. Schönherr, H. & Cernak, T. *Angew. Chem. Int. Edn* **52**, 12256–12267 (2013).
5. Moir, M., Danon, J. J., Reekie, T. A. & Kassiou, M. *Expert Opin. Drug Discov.* **14**, 1127–1149 (2019).
6. Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. *Chem. Soc. Rev.* **45**, 546–576 (2019).
7. Sono, M., Roach, M. P., Coulter, E. D. & Dawson, J. H. *Chem. Rev.* **96**, 2841–2888 (1996).
8. Denisov, I. G., Makris, T. M., Sligar, S. G. & Schlichting, I. *Chem. Rev.* **105**, 2253–2278 (2005).
9. Chen, M. S. & White, C. *Science* **318**, 783–787 (2007).
10. Zhao, J., Nanjo, T., de Lucca, E. C. & White, M. C. *Nature Chem.* **11**, 213–221 (2019).

Microbiology

Gut pain sensors help to combat infection

Romana R. Gerner & Manuela Raffatellu

The mammalian gut must defend against a variety of infectious agents. Neurons, cells not usually thought of as first-responders during infection, are now found to aid the gut's barrier function and stop bacteria from spreading elsewhere.

The recognition and neutralization of harmful bacteria in the gut is generally thought to be orchestrated by epithelial and immune cells. Writing in *Cell*, Lai *et al.*[1] report that a subset of gut neurons also have an unexpected crucial role in the intestinal response to infection.

The gut is exposed to food, antigens (molecules that can trigger an immune response if they are recognized as 'non-self'), resident microbes that are normally harmless (commensal microbes) and harmful microbes (pathogens). Thus, gut cells have the difficult task of discerning friend from foe. Cooperative interactions between epithelial cells and immune cells are key to managing this complex balancing act, coordinating tolerance to food antigens and to commensal microbes, but initiating protective immune responses against pathogens. Nerve cells in the gut (comprising the enteric nervous system) can sense microbe-derived molecules, and these neurons interact with epithelial cells and immune cells to promote defence responses against microbes[2].

The human enteric nervous system encompasses a complex network of an estimated 10^8 neurons, which are essential for regulating many gut functions, including blood flow and the movement of the intestine's contents[3]. Moreover, certain subsets of these nerve cells interact closely with and modulate components of the gut's immune system[4]. One such type of neuron is called a nociceptor. Nociceptors elicit the perception of pain or discomfort in response to potentially harmful stimuli such as intense heat and cold, reactive chemicals or mechanical injury[5]. They also directly detect pathogens and pathogen-produced molecules, which can evoke a sensation of pain during infection[6]. Whether nociceptors contribute directly to impeding bacterial invasion of host tissues has been a matter of speculation.

Salmonella is a pathogenic bacterium that is a frequent cause of food-borne illness. It can trigger a variety of conditions, ranging from inflammatory diarrhoea (gastroenteritis) to life-threatening complications in situations in which infection spreads beyond the gut to other sites in the body[7,8]. When *Salmonella* reaches the gut, one of the major sites of tissue invasion are the dome-shaped follicles called Peyer's patches (Fig. 1). As key sensors of the gut that aid immune defences, these follicles use immune cells and specialized epithelial cells called M cells to monitor and respond to pathogens and commensal microbes. M cells can take up antigens from the lumen of the gut and transfer them to underlying immune cells[9], which then either initiate a protective immune response to the antigen or tolerate the antigen's presence.

Although Peyer's patches are important for monitoring the contents of the intestine, certain pathogenic agents, including *Salmonella*, norovirus (a common viral cause of gastroenteritis) and prions (infectious, disease-causing proteins), exploit M cells as sites of tissue entry[10,11]. Notably, although Peyer's patches are adjacent to neurons, including nociceptors, the functional consequences of

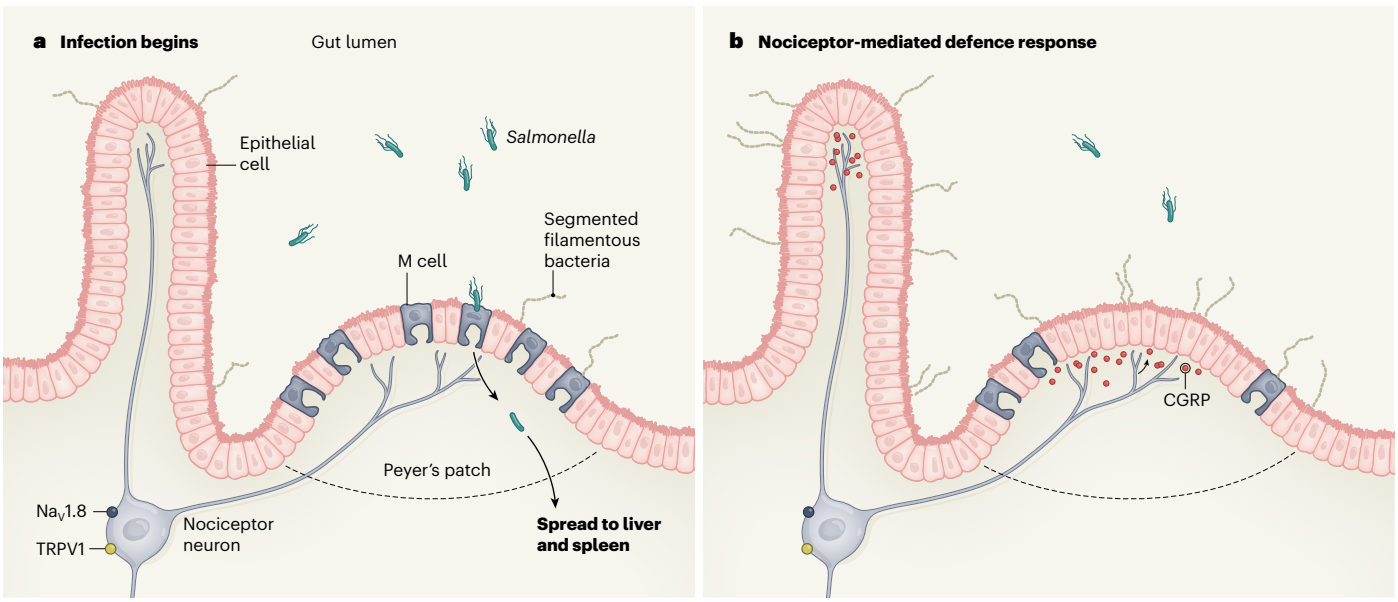


Figure 1 | Neurons in the mammalian gut help to defend against bacterial infection. Lai *et al.*[1] report that a type of gut neuron, called a nociceptor, which expresses the proteins TRPV1 and $Na_V1.8$, can thwart infection by *Salmonella* bacteria in mice. **a**, *Salmonella* can invade gut tissue by leaving the gut lumen to enter M cells — specialized epithelial cells in a region called a Peyer's patch. These microbes can then spread elsewhere. Segmented filamentous bacteria (SFB), which are normally present in the gut and can bind to M cells or other gut epithelial cells, can help to limit *Salmonella* infection[13]. **b**, The authors report that these nociceptors orchestrate a defence response to *Salmonella* infection, the hallmarks of which include a decrease in the number of M cells and an increase in SFB colonization of the gut. These changes were associated with a decrease in *Salmonella* infection and diminished spread of the bacteria beyond the gut. Lai *et al.* report that nociceptor secretion of the neuropeptide CGRP regulates the number of M cells, along with SFB levels.

this close proximity were not fully understood previously.

Lai *et al*. assessed the role of nociceptors in the gut of mice infected with the pathogen *Salmonella enterica* serovar Typhimurium. The authors report that the presence of a subset of gut nociceptors (specifically, those that express the ion-channel proteins TRPV1 and $Na_V1.8$) protect the gut against invasion by *Salmonella* and the subsequent spread of this bacterium to sites such as the liver and spleen. Intriguingly, the authors found that the protective effects of nociceptors were not mediated by well-known antimicrobial defence mechanisms, such as activation of immune cells or alterations in the levels of antimicrobial peptides that are produced by gut cells. Instead, during infection with *Salmonella*, these nociceptors orchestrated a reduction in the number of M cells. Because M cells are a key entry point for *Salmonella*, this reduction would probably have the consequence of reducing the surface area available for *Salmonella* to invade.

The authors analysed the composition of gut bacteria in the absence of *Salmonella* infection, using mice with gut nociceptors that were genetically engineered to lack either TRPV1 or $Na_V1.8$ channel proteins. Compared with animals that expressed these proteins, both types of engineered mouse had lower levels of segmented filamentous bacteria (SFB), a group of commensal microbes that attach to gut epithelial cells, and particularly to M cells[12]. Such commensal bacteria are crucial for providing resistance against gut colonization by pathogens, including *Salmonella*[13].

Lai and colleagues investigated whether there was a connection between a decrease in M cells and the extent of SFB colonization of the Peyer's patches. The authors demonstrated that M-cell depletion mediated by nociceptors, or triggered through an antibody-mediated experimental approach, led to an increase in this colonization, suggesting that the number of M cells can modulate SFB colonization in the gut (although the exact mechanisms responsible were not fully determined). This outcome was beneficial because it limited *Salmonella* infection, presumably because the higher presence of SFB and the depletion of M cells together resulted in a reduction of invasion sites available for *Salmonella*. Finally, Lai *et al*. report that when TRPV1-expressing nociceptors encountered *Salmonella*, the neurons released a neuropeptide called CGRP. This small molecule enables communication between cells. CGRP was directly able to regulate M-cell abundance and function, as well as to regulate SFB levels in the gut.

The authors have uncovered a previously unrecognized role for nociceptors in host defence against *Salmonella* infection. These remarkable findings reveal a complex loop of interactions between epithelial cells, neurons and microbes in the mammalian gut, adding another layer of complexity to our understanding of gut immunity. Whether nociceptor-mediated responses help to defend against a variety of other microbial pathogens remains to be determined. Indeed, nociceptors have been reported to protect mice during infection by the bacterial pathogen *Citrobacter rodentium*[14].

A key area for future investigation will be to determine whether Lai and colleagues' findings have relevance for human health. For example, one area that would be worth studying is whether long-term use of pain-blocking opioid drugs, such as morphine, might affect nociceptor-mediated antibacterial defence. This is of interest because nociceptors are the main target of opioids, and administering morphine to mice changes the gut's microbial composition[15,16]. Moreover, morphine use promotes the spread of certain types of microbe (Gram-negative bacteria) from the gut to elsewhere in the body, a process that can lead to sepsis, a potentially life-threatening immune response to infection[15,16]. Future research that explores interactions between neurons and immune cells during infection could uncover further exciting findings that will profoundly influence our understanding of host defence.

"These remarkable findings reveal a complex loop of interactions between epithelial cells, neurons and microbes."

Romana R. Gerner and **Manuela Raffatellu** are in the Department of Pediatrics, Division of Host–Microbe Systems and Therapeutics, University of California, San Diego, La Jolla, California 92093, USA.
e-mails: rgerner@health.ucsd.edu; manuelar@health.ucsd.edu.

1. Lai, N. Y. *et al*. *Cell* **180**, 33–49 (2020).
2. Yoo, B. B. & Mazmanian, S. K. *Immunity* **46**, 910–926 (2017).
3. Kulkarni, S. *et al*. *J. Neurosci.* **38**, 9346–9354 (2018).
4. Schneider, S., Wright, C. M. & Heuckeroth, R. O. *Annu. Rev. Physiol.* **81**, 235–259 (2019).
5. Julius, D. & Basbaum, A. I. *Nature* **413**, 203–210 (2001).
6. Baral, P., Udit, S. & Chiu, I. M. *Nature Rev. Immunol.* **19**, 433–447 (2019).
7. Majowicz, S. E. *et al*. *Clin. Infect. Dis.* **50**, 882–889 (2010).
8. Gordon, M. A. *J. Infect.* **56**, 413–422 (2008).
9. Mabbott, N. A., Donaldson, D. S., Ohno, H., Williams, I. R. & Mahajan, A. *Mucosal Immunol.* **6**, 666–677 (2013).
10. Jung, C., Hugot, J.-P. & Barreau, F. *Int. J. Inflam.* **2010**, 823710 (2010).
11. Chiocchetti, R. *et al*. *Cell Tissue Res.* **332**, 185–194 (2008).
12. Meyerholz, D. K., Stabel, T. J. & Cheville, N. F. *Infect. Immun.* **70**, 3277–3280 (2002).
13. Garland, C. D., Lee, A. & Dickson, M. R. *Microb. Ecol.* **8**, 181–190 (1982).
14. Ramirez, V. T. *et al*. *J. Infect. Dis.* https://doi.org/10.1093/infdis/jiaa014 (2020).
15. Wang, F. *et al*. *Sci. Rep.* **8**, 3596 (2018).
16. Hilburger, M. E. *et al*. *J. Infect. Dis.* **176**, 183–188 (1997).

This article was published online on 20 April 2020.

Cancer genetics

Not all driver mutations are equal

Victoria L. Bae-Jump & Douglas A. Levine

A study of cancer-associated mutations in normal endometrial glands of the uterus has now been performed using whole-genome sequencing. The analysis sheds light on the early changes that lead to invasive disease. **See p.640**

Understanding how normal tissues give rise to cancer is crucial for improving prevention and early detection of this deadly disease. Over the past two decades, the genomic profiles of most types of invasive cancer have been catalogued; however, similar profiling of normal tissues presents a unique set of challenges. Cancer tissues are often abundantly available from biopsies or surgery, but samples from normal tissues tend to be much smaller, and specimen-collection practices are less well established, making it hard to gather high-quality material. Moore *et al*.[1] overcome these challenges on page 640, and successfully catalogue cancer-driving mutations in normal endometrial glands.

Endometrial glands are abundant in the lining of the uterus, where they secrete hormones and other substances that are essential for normal menstruation and embryonic development. Endometrial cancer is the sixth most common cancer in women worldwide, with more than 382,000 cases annually[2]. The mortality rate has increased over the past decade[3], heightening the need for prevention and early detection of this disease.

Moore *et al.* obtained 257 normal endometrial glands from 28 women of various ages. In each case, the authors meticulously isolated the glands using a technique called laser-capture microdissection to separate the epithelial tissue, which lines the gland, from the surrounding stromal cells that make up the gland's connective tissue. They then performed whole-genome sequencing of the epithelial samples and various other normal tissues from the same women, using a protocol they had developed that is tailored to the analysis of small amounts of DNA.

The group analysed these sequences to identify mutations that are unique to the normal endometrial glands, as well as endometrium-specific changes in the number of copies of any genetic region (caused by duplication or deletion of DNA). They found that, in almost 90% of individuals, the normal endometrial tissue contained driver mutations — which give cells a selective advantage over non-mutated counterparts, and so are thought to promote cancer development. Nearly 60% of the endometrial glands in these women contained one or more drivers.

The authors found 12 genes that contained driver mutations with statistically increased prevalence in normal endometrial tissue compared with that in other tissues. These genes are all known to be frequently mutated in cancer, and, collectively, these mutations have the potential to affect many cellular processes. However, isolated mutations in the individual genes, as was typically the case in Moore and colleagues' samples, are probably insufficient to make a tissue become cancerous[4].

A remarkable finding is that each endometrial gland seems to be clonal — that is, all the cells in the gland are derived from a single epithelial progenitor cell. It might be expected that each gland could develop multiple independent mutations, but the authors' discovery of clonality indicates that there is instead a uniformity to the mutational process.

As would be expected, the number of mutations increased with age, at the rate of about 29 nucleotide substitutions per gland per year during adult life. Moore *et al.* reconstructed the phylogeny (the evolutionary development and diversification) of individual glands to document the initial presentation and spread of driver mutations through the tissue over time. They report that many glands that were located in close physical proximity in the uterine wall displayed distant phylogeny. This suggests that the cellular populations in each gland remain genetically isolated, providing many separate opportunities for cancer to develop. The researchers also provide evidence that driver mutations can arise at any time, occurring in some women in their first decade of life and in others at various stages of adulthood. This insight is important because the typical timeline between developing driver mutations and cancer is not yet well defined.

The group's rigorous methods for sample isolation and sequencing, coupled with their well-developed bioinformatics algorithms, mean that the results of this study should be highly reliable and reproducible. But one caveat is that the authors isolated endometrial glands from a select population of women: most samples were obtained from people undergoing evaluation for infertility, from organ donors, or from women who had died of non-gynaecological causes. Both infertility and nulliparity (having never given birth) are known independent risk factors for endometrial cancer[5]. And samples collected from women who had died of non-gynaecological causes might be more likely than the average endometrial gland to contain low-risk driver mutations that have less potential to trigger cancer, given that these women died without having developed endometrial cancer.

Future studies would benefit from a more-representative cross-sectional population. The inclusion of women who have conditions that are well-known precursors to cancer, such as atypical endometrial hyperplasia (in which the lining of the uterus becomes abnormally thick) could help in this regard. Researchers might then be able to define a robust landscape of changes that occur during the progression from normal to precancerous tissue to invasive disease. This approach might also help to define the pathogenicity of, and possible necessity for, individual driver mutations that lead to the development of cancer.

"The authors' findings should be useful for ongoing research to detect endometrial cancer at early stages."

Another caveat is the discrepancy between driver mutations identified by Moore *et al.* and those from other cancer-genome projects, including The Cancer Genome Atlas[6]. Although the most frequently mutated genes identified in the current study have been previously reported in endometrial cancers, several of the most commonly mutated genes in this cancer are notably not mutated in Moore and colleagues' samples. The group found mutations in these well-known drivers in less than 2% of the normal endometrial glands that they studied — a surprisingly low frequency, because one would expect that the drivers present in all cancer cells would be the first to arise in normal tissue. This discrepancy probably hints at unknown aspects of the multistep process of tumour initiation, in which certain mutations must arise before others. Determining when and how gatekeeper mutations occur and permit the next step in tumour development will require further analyses of benign, premalignant and invasive tissues.

Knowing that the compilation of driver mutations in normal endometrial glands is different from those found in established endometrial cancers might change the approach for further research into the prevention and early detection of this disease. Determining the role of these mutations in concert with other known risk factors, such as nulliparity, obesity, race and genetic predisposition, will help to better identify women who are at risk of endometrial cancer. Even before we obtain this information, Moore and colleagues' findings should be useful for ongoing research to detect endometrial cancer at early stages, which includes analyses of cell-free DNA circulating in blood, tampon-based collection of vaginal secretions and liquid-based examination of cervical tissues[7–9]. More broadly, a better overall understanding of the normal mutational spectra in tissues will refine our knowledge of the consequences of specific cancer drivers for many solid tumours.

Victoria L. Bae-Jump is at Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. **Douglas A. Levine** is at Perlmutter Cancer Center, NYU Langone Health, New York, New York 10016, USA.
e-mails: victoria.bae-jump@unchealth.unc.edu; douglas.levine@nyulangone.org

1. Moore, L. *et al. Nature* **580**, 640–646 (2020).
2. Bray, F. *et al. CA Cancer J. Clin.* **68**, 394–424 (2018).
3. Siegel, R. L., Miller, K. D. & Jemal, A. *CA Cancer J. Clin.* **70**, 7–30 (2020).
4. Joshi, A., Miller, C. Jr., Baker, S. J. & Ellenson, L. H. *Am. J. Pathol.* **185**, 1104–1113 (2015).
5. Yang, H. P. *et al. Br. J. Cancer* **112**, 925–933 (2015).
6. Levine, D. A. & The Cancer Genome Atlas Research Network. *Nature* **497**, 67–73 (2013).
7. Sangtani, A. *et al. Gynecol. Oncol.* **156**, 387–392 (2020).
8. Wang, Y. *et al. Sci. Transl. Med.* **10**, eaap8793 (2018).
9. Wan, J. C. M. *et al. Nature Rev. Cancer* **17**, 223–238 (2017).

This article was published online on 22 April 2020.

Article

Nightside condensation of iron in an ultrahot giant exoplanet

https://doi.org/10.1038/s41586-020-2107-1

Received: 11 September 2019

Accepted: 24 January 2020

Published online: 11 March 2020

Check for updates

David Ehrenreich[1]✉, Christophe Lovis[1], Romain Allart[1], María Rosa Zapatero Osorio[2], Francesco Pepe[1], Stefano Cristiani[3], Rafael Rebolo[4], Nuno C. Santos[5,6], Francesco Borsa[7], Olivier Demangeon[5], Xavier Dumusque[1], Jonay I. González Hernández[4], Núria Casasayas-Barris[4], Damien Ségransan[1], Sérgio Sousa[5], Manuel Abreu[8,9], Vardan Adibekyan[5], Michael Affolter[10], Carlos Allende Prieto[4], Yann Alibert[10], Matteo Aliverti[7], David Alves[8,9], Manuel Amate[4], Gerardo Avila[11], Veronica Baldini[3], Timothy Bandy[10], Willy Benz[10], Andrea Bianco[7], Émeline Bolmont[1], François Bouchy[1], Vincent Bourrier[1], Christopher Broeg[10], Alexandre Cabral[8,9], Giorgio Calderone[3], Enric Pallé[4], H. M. Cegla[1], Roberto Cirami[3], João M. P. Coelho[8,9], Paolo Conconi[7], Igor Coretti[3], Claudio Cumani[11], Guido Cupani[3], Hans Dekker[11], Bernard Delabre[11], Sebastian Deiries[11], Valentina D'Odorico[3,12], Paolo Di Marcantonio[3], Pedro Figueira[5,13], Ana Fragoso[4], Ludovic Genolet[1], Matteo Genoni[7], Ricardo Génova Santos[4], Nathan Hara[1], Ian Hughes[1], Olaf Iwert[11], Florian Kerber[11], Jens Knudstrup[11], Marco Landoni[7], Baptiste Lavie[1], Jean-Louis Lizon[11], Monika Lendl[1,14], Gaspare Lo Curto[13], Charles Maire[1], Antonio Manescau[11], C. J. A. P. Martins[5,15], Denis Mégevand[1], Andrea Mehner[13], Giusi Micela[16], Andrea Modigliani[11], Paolo Molaro[3,17], Manuel Monteiro[5], Mario Monteiro[5,6], Manuele Moschetti[7], Eric Müller[11], Nelson Nunes[8], Luca Oggioni[7], António Oliveira[8,9], Giorgio Pariani[7], Luca Pasquini[11], Ennio Poretti[7,18], José Luis Rasilla[4], Edoardo Redaelli[7], Marco Riva[7], Samuel Santana Tschudi[13], Paolo Santin[3], Pedro Santos[8,9], Alex Segovia Milla[1], Julia V. Seidel[1], Danuta Sosnowska[1], Alessandro Sozzetti[19], Paolo Spanò[7], Alejandro Suárez Mascareño[4], Hugo Tabernero[2,5], Fabio Tenegi[4], Stéphane Udry[1], Alessio Zanutta[7] & Filippo Zerbi[7]

Ultrahot giant exoplanets receive thousands of times Earth's insolation[1,2]. Their high-temperature atmospheres (greater than 2,000 kelvin) are ideal laboratories for studying extreme planetary climates and chemistry[3–5]. Daysides are predicted to be cloud-free, dominated by atomic species[6] and much hotter than nightsides[5,7,8]. Atoms are expected to recombine into molecules over the nightside[9], resulting in different day and night chemistries. Although metallic elements and a large temperature contrast have been observed[10–14], no chemical gradient has been measured across the surface of such an exoplanet. Different atmospheric chemistry between the day-to-night ('evening') and night-to-day ('morning') terminators could, however, be revealed as an asymmetric absorption signature during transit[4,7,15]. Here we report the detection of an asymmetric atmospheric signature in the ultrahot exoplanet WASP-76b. We spectrally and temporally resolve this signature using a combination of high-dispersion spectroscopy with a large photon-collecting area. The absorption signal, attributed to neutral iron, is blueshifted by −11 ± 0.7 kilometres per second on the trailing limb, which can be explained by a combination of planetary rotation and wind blowing from the hot dayside[16]. In contrast, no signal arises from the nightside close to the morning terminator, showing that atomic iron is not absorbing starlight there. We conclude that iron must therefore condense during its journey across the nightside.

Two transits of the short-period (1.81 d) giant exoplanet WASP-76b[17–19] were observed on 2 September 2018 (epoch 1) and 30 October 2018 (epoch 2) with the Echelle Spectrograph for Rocky Exoplanets and Stable Spectroscopic Observations (ESPRESSO) at the European Southern Observatory Very Large Telescope (VLT) located on Cerro Paranal, Chile (see the observation log in Methods and Extended Data Fig. 1). ESPRESSO is a fibre-fed, stabilized and high-resolution spectrograph[20] able to collect the light from any combination of the four VLT 8-m Unit Telescopes (UTs). During each epoch, we acquired data with UT3 only. We used the single high-resolution 2 × 1-binning mode (average spectral resolution of 138,000) and exposure times of 600 s and 300 s with the slow read-out mode to record 35 and 70 spectra of the bright (V = 9.5),

A list of affiliations appears at the end of the paper.

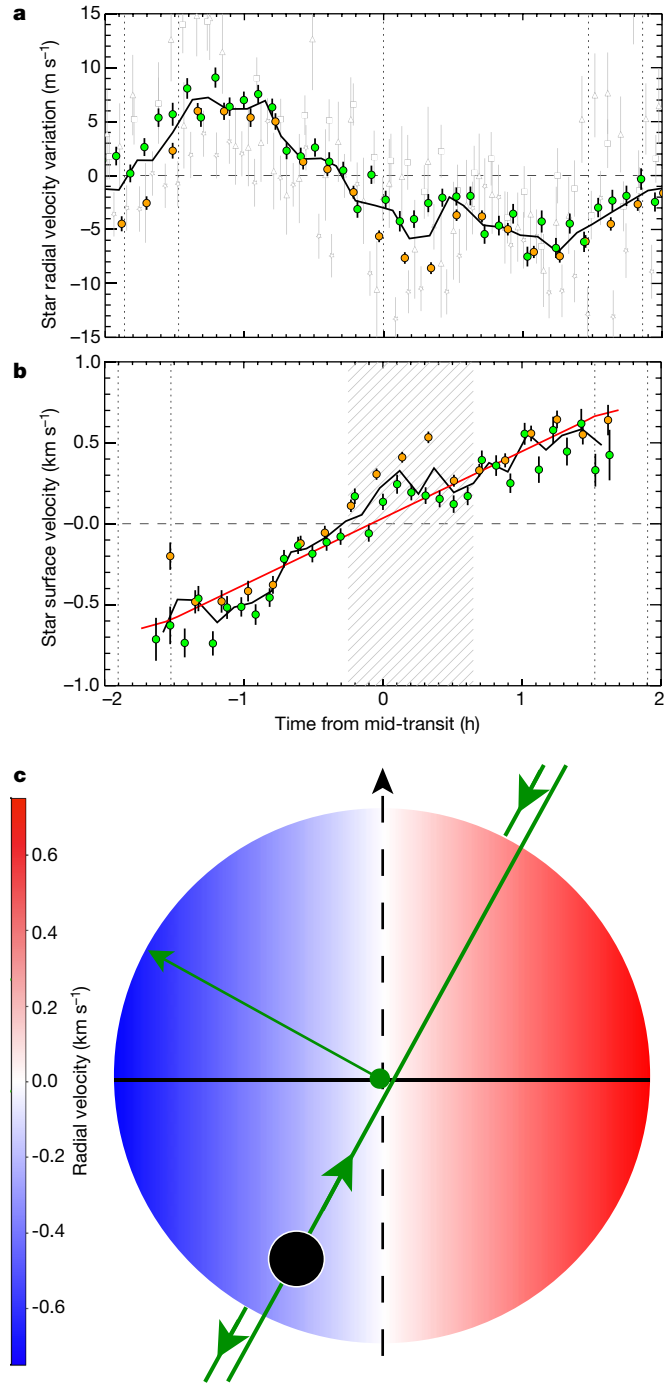


Fig. 1 | Rossiter–McLaughlin effect of WASP-76b. a, 'Classical' analysis of the effect, showing the radial velocities integrated over the whole stellar disk for ESPRESSO epoch 1 (orange data points with error bars), epoch 2 (green data points with error bars), both epochs combined (black thick curve) and three previous transits observed with HARPS (grey data points with error bars[19]). All error bars represent 1σ uncertainties. **b**, 'Reloaded' analysis of the effect showing the stellar surface velocities behind the disk of the planet (epoch 1 in orange, epoch 2 in green). The red curve is a fit with a stellar surface model assuming solid-body rotation. Vertical dotted lines indicate the transit contacts and mid-time. The hatched area delimits the times when the planetary absorption signal crosses the Doppler shadow. The 1σ uncertainties (error bars) have been propagated accordingly from the errors calculated by the ESPRESSO pipeline. Velocity scales are in the stellar rest frame. **c**, Sketch of the WASP-76 system (to scale) as seen from Earth. Arrows show the projected spin axes of the planetary orbit (green) and the star (black).

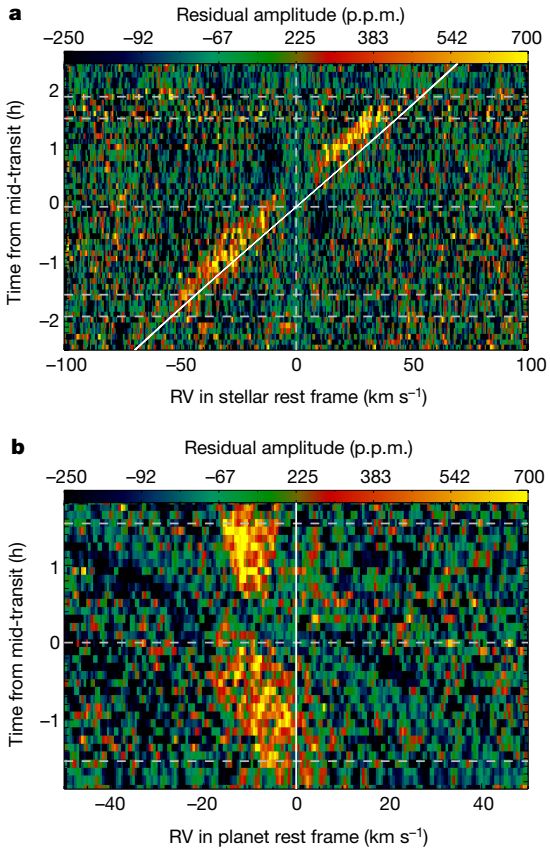


Fig. 2 | Planetary absorption signature in the residuals of the Doppler shadow subtraction. a, In the stellar rest frame, the planetary absorption signal (sloping orange feature) appears close to the expected Keplerian of the planet, superimposed in white with its 1σ uncertainty (barely visible). Transit contacts are shown by white horizontal dashed lines. The gap around 0 km s^{-1} corresponds to the position of the Doppler shadow before its subtraction. **b**, In the planet rest frame, the planetary absorption signal is asymmetric and progressively blueshifts after ingress.

F7 star WASP-76 during epochs 1 and 2, respectively. The ESPRESSO pipeline version 1.3.2 was used to produce one-dimensional (1D) spectra and cross-correlation functions (CCFs). The CCFs, which are average stellar line profiles, were extracted using an F9 mask over a velocity range of [−150, +150] km s^{-1}, with a step of 0.5 km s^{-1}. This stellar mask contains 4,653 spectral lines in the wavelength range between 380 nm and 788 nm covered by ESPRESSO; most of the spectroscopic information in the mask is contained in electronic transitions of neutral iron (Fe; see Methods).

We used the spectra to revise the stellar parameters and the CCF peak position to monitor the radial velocities of the integrated stellar disk during the transit of the planet (see Methods and Extended Data Table 1). The planet blocks different parts of the rotating stellar surface during the transit, resulting in a spectroscopic anomaly known for eclipsing binaries and exoplanets as the Rossiter–McLaughlin effect (see, for example, ref. [21]). The shape of the anomaly observed for WASP-76b (Fig. 1a) shows that the planet orbit is prograde, and its orbital spin is approximately aligned with the rotational spin of the star. For this bright star, ESPRESSO yields an average photon-noise-limited precision of 70 cm s^{-1} and 85 cm s^{-1} per 10 min and 5 min exposure in epochs 1 and 2, respectively. This precision is high enough to reveal a small 'bump' towards positive radial velocities occurring 30 min after mid-transit during each epoch. We could also find it a posteriori in previous data taken with the HARPS spectrograph (Fig. 1a).

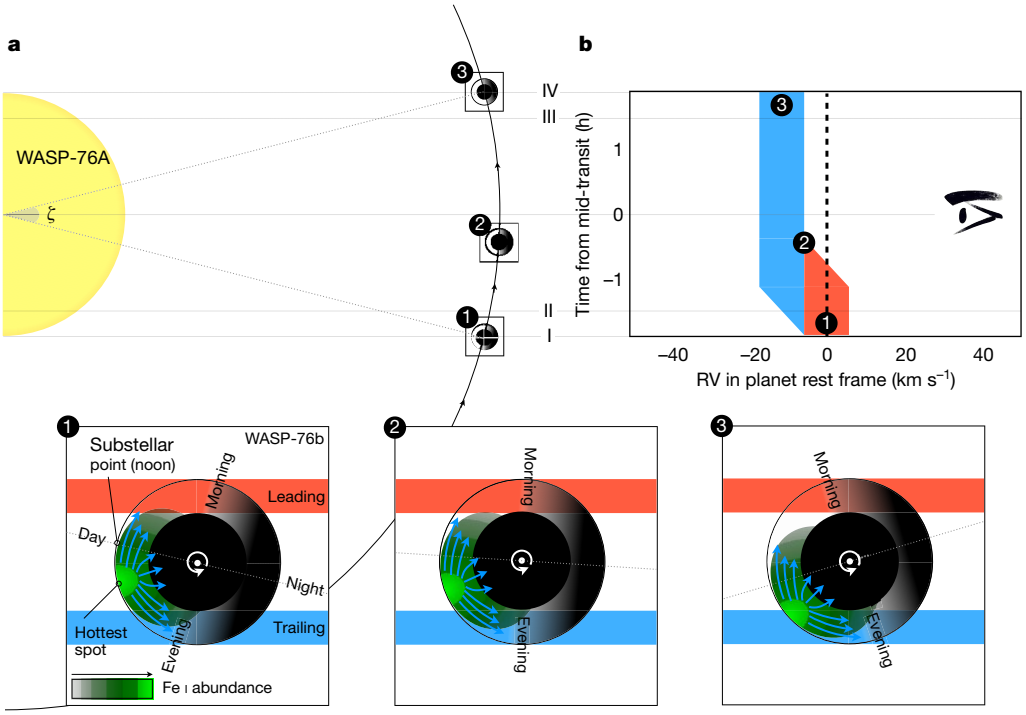


Fig. 3 | Polar view of the WASP-76 system. a, The star WASP-76A and planet WASP-76b are represented to scale in size and distance. The planet is shown at different transit stages, with the transit contacts I, II, III and IV. During transit, the angle ζ between the planet terminator and the line of sight (dashed line in the middle) changes by $2\arcsin(R_*/a) = 29.4°$, where a is the semi-major axis. **b**, Sketch of the absorption signature observed during transit, in the planet rest frame. The numbers refer to the insets across the bottom of the figure. Inset 1: during ingress, iron on the dayside is visible through the leading limb and creates an absorption around 0 km s^{-1} due to the combination of planetary rotation and day-to-night winds (blue arrows). The trailing limb enters the stellar disk and progressively blueshifts the signal. Inset 2: the signal around 0 km s^{-1} disappears as soon as no more iron is visible in the leading limb. Only the trailing limb contributes to the signal, which remains blueshifted around −11 km s^{-1}. Inset 3: the signal remains at this blueshifted velocity until the end of the transit.

We retrieved the CCFs of the stellar surface occulted by the planet during the transit[22] (the 'local' CCFs). These local CCFs, shifted to the stellar rest frame, are displayed for individual epochs in Extended Data Fig. 4. They exhibit a dark slanted feature called the Doppler shadow[1]. The radial velocity of the Doppler shadow across the transit is related to the stellar projected rotational velocity $v_{eq}\sin i_*$ (where i_* is the stellar inclination and v_{eq} is the equatorial rotation velocity of the star) and the projected spin–orbit obliquity λ (ref. [22]). For each in-transit exposure, we fitted the Doppler shadow with a Voigt profile. The radial velocity of the peak corresponds to the local velocity of the stellar surface behind the planet (see Methods and Fig. 1b). A fit to the data with a stellar surface model described in Methods yields $\lambda = 61° \pm 7°$ and a slow projected stellar rotation ($v_{eq}\sin i_* = 1.5 \pm 0.3$ km s^{-1}). The system geometry is sketched in Fig. 1c.

We removed the Doppler shadow by subtracting the Voigtian fits to the data for each exposure and searched for faint planetary signals in the residuals. Any signal tied to the planet should move with a velocity close to the planet Keplerian velocity, which varies between −53 km s^{-1} and +53 km s^{-1} during transit. The residual maps, shown for both epochs combined and each epoch separately in Fig. 2a and Extended Data Fig. 5, respectively, show a slanted residual absorption signature close to the expected radial velocity of the planet. In contrast to the negative Doppler shadow, the slanted signature appears as a positive signal. The gap between −0.2 h and +0.7 h around mid-transit results from the subtraction of the Doppler shadow.

The absorption signal arises from the cross-correlation of the planetary atmosphere transmission spectrum with the stellar mask dominated by atomic iron lines. Therefore, it traces the presence of atomic iron in the atmosphere of WASP-76b, as found in the atmospheres of other ultrahot gas giants[10–12,23]. The planetary absorption overlaps the Doppler shadow at the same transit phase as the 'bump' in Fig. 1a, meaning that this anomaly in the Rossiter–McLaughlin effect is due to the presence of a hot planetary atmosphere, as described in previous works[23,24]. We excluded this phase range from our analysis.

We studied the absorption signature in the planetary rest frame (Fig. 2b), using our newly derived orbital solution (see Methods). It appears asymmetric and mostly blueshifted. We fitted the absorption signal with Gaussians to retrieve its amplitude, radial velocity and full-width at half-maximum (FWHM) as a function of time (see Extended Data Fig. 7).

The radial velocity of the planetary signature (Extended Data Fig. 7a) is slightly blueshifted (between 0 km s^{-1} and −5 km s^{-1}) at ingress. It progressively blueshifts down to about -11 ± 0.7 km s^{-1}, reached at −0.4 h from mid-transit. It remains blueshifted until the end of transit. The amplitude can be converted to a differential transit depth δ_{atm} caused by atmospheric absorption (Extended Data Fig. 7c). We measure a mean absorption signal of 494 ± 27 ppm (weighted by the uncertainties), which corresponds to absorption by about 1.8 atmospheric scale heights calculated assuming a dayside temperature[25] of 2,693 K. During the first half of the transit, the signal contrast is 434 ± 32 ppm close to the planet rest velocity; it increases (with a significance of 3.5σ) after +1 h from mid-transit, up to 628 ± 49 ppm towards the end of the transit, where the signal is significantly blueshifted (16σ; Extended Data Fig. 7a). An atmospheric signal centred at non-zero radial velocities indicates a motion of the absorber in the planet rest frame, typically due to winds[16]. The blueshifted signal is highly significant compared to the absence of a signal (0 ± 49 ppm) observed at the same time around 0 km s^{-1} at the 9σ confidence level ($628\ \text{ppm}/(49\ \text{ppm} \times \sqrt{2}) \approx 9\sigma$). Finally, the FWHM (Extended Data Fig. 6b) has a weighted-mean value of

8.6 ± 0.7 km s^{-1}. This width could result from the combination of the tidally-locked planetary rotation (5.3 km s^{-1}), thermal broadening (about 0.7 km s^{-1}) and turbulent motions due to winds.

Three-dimensional global climate models will be needed to take full advantage of these spectrally and temporally resolved figures. Meanwhile, we can craft a 'toy model' to understand qualitatively the temporal evolution of the atmospheric signature. Ultrahot gas giants have dayside temperatures commensurate with the surface of cool stars. There, most molecules should thermally dissociate, resulting in a composition dominated by atoms and ions. These partially ionized atmospheres could give rise to frictional drag caused by Lorentz forces, slowing down the characteristic timescale for heat advection[7,8]. Consequently, heat redistribution to the nightside should be relatively inefficient and the day–night temperature contrast correspondingly large. This could also result in day-to-night, longitudinally symmetric winds[14,16] and recombination and dissociation of atoms and molecules at the evening and morning terminators[9], respectively. Our model is sketched in Fig. 3 and makes use of the following ingredients. (i) A tidally locked rotation redshifting and blueshifting the signal at the leading and trailing limb, respectively, by ±5.3 km s^{-1}. (ii) An absorber (neutral iron) found in the hottest part of the planet dayside (the 'hotspot') and absent from the colder nightside. (iii) A longitudinal offset of the hotspot towards the evening terminator (this is in tension with the existence of strong drag forces but is necessary to explain the asymmetry between the beginning and end of transit). (iv) A uniform day-to-night wind, previously observed for HD 209458b[16]. We assumed the day-to-night wind imprints a −5.3 km s^{-1} blueshift at both limbs, therefore compensating the planetary rotation redshift at the leading limb (shifting the signal towards 0 km s^{-1}) while increasing the blueshift at the trailing limb to −10.6 km s^{-1}. This wind speed lies at the upper bound of the range expected by theory[7,15,26,27]; however, most existing studies have focused on planets cooler than WASP-76b. Finally, we considered (v) the variation of the angle $\zeta = 2\arcsin(R_*/a)$ between the planet terminator and the line of sight, where a is the semi-major axis of the planet. Owing to the tight orbit of WASP-76b, this angle varies by 29.4° during transit. This effect brings the region containing atomic iron in and out of view during the transit.

The transit of our model in front of the star follows the three-step scenario depicted in Fig. 3. (1) At ingress, only the leading limb contributes to the signal; there, the line of sight crosses a fraction of the dayside containing iron atoms that absorb the starlight between 0 km s^{-1} and −5 km s^{-1}. The trailing limb, entering the stellar disk, also contains iron atoms that start to blueshift the signal. (2) As soon as ζ is large enough to take the patch of absorbing iron atoms out of view from the leading limb, its contribution disappears. Only the trailing limb now contributes to the signal; the planetary rotation plus day-to-night winds bring it to −10.6 km s^{-1}. Planetary rotation could still broaden the signal after the disappearance of the leading limb signal because an absorption crescent on the trailing limb still features differential rotation from the pole to the equator. The signal then keeps a constant blueshift until egress (3) while its absorption depth increases as the hotspot comes into view in the trailing limb. We can quantify the temperature increase between the evening terminator and the hotspot by considering that $\delta_{\mathrm{atm}}^{\mathrm{day}}/\delta_{\mathrm{atm}}^{\mathrm{eve}} = T^{\mathrm{day}}/T^{\mathrm{eve}}$ (see Methods). Given the values quoted above, the temperature must increase by a factor of 1.5 ± 0.2 across the evening terminator towards the dayside. This is consistent with expectations for a strong day–night temperature contrast for ultrahot gas giants[28] and yields an evening temperature of $1{,}795 \pm 242$ K. Meanwhile, the egress of the leading limb does not have any apparent impact, confirming its lack of contribution.

We conclude from this that neutral iron atoms must be present on the dayside and the evening terminator, but are much less abundant or even absent from the nightside and the morning terminator. Therefore, iron must condense across the nightside. Nightside clouds have been suggested from thermal phase curves of hot gas giants[28,29]. On WASP-76b and similarly hot planets, these clouds could be made out of iron droplets, because liquid iron is the most stable high-temperature iron-bearing condensate[30]. Hence, it could literally rain iron on the nightside of WASP-76b.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2107-1.

1. Collier-Cameron, A. et al. Line-profile tomography of exoplanet transits - II. A gas-giant planet transiting a rapidly rotating A5 star. *Mon. Not. R. Astron. Soc.* **407**, 507–514 (2010).
2. Gaudi, B. S. et al. A giant planet undergoing extreme-ultraviolet irradiation by its hot massive-star host. *Nature* **546**, 514–518 (2017).
3. Evans, T. M. et al. An ultrahot gas-giant exoplanet with a stratosphere. *Nature* **548**, 58–61 (2017).
4. Parmentier, V. et al. From thermal dissociation to condensation in the atmospheres of ultra hot Jupiters: WASP-121b in context. *Astron. Astrophys.* **617**, A110 (2018).
5. Lothringer, J. D., Barman, T. & Koskinen, T. Extremely irradiated hot Jupiters: non-oxide inversions, H^- opacity, and thermal dissociation of molecules. *Astrophys. J.* **866**, 27 (2018).
6. Kitzmann, D. et al. The peculiar atmospheric chemistry of KELT-9b. *Astrophys. J.* **863**, 183 (2018).
7. Miller-Ricci Kempton, E. & Rauscher, E. Constraining high-speed winds in exoplanet atmospheres through observations of anomalous Doppler shifts during transit. *Astrophys. J.* **751**, 117 (2012).
8. Komacek, T. D. & Showman, A. P. Atmospheric circulation of hot Jupiters: dayside-nightside temperature differences. *Astrophys. J.* **821**, 16 (2016).
9. Bell, T. J. & Cowan, N. B. Increased heat transport in ultra-hot Jupiter atmospheres through H_2 dissociation and recombination. *Astrophys. J.* **857**, L20 (2018).
10. Fossati, L. et al. Metals in the exosphere of the highly irradiated planet WASP-12b. *Astrophys. J.* **714**, L222 (2010).
11. Hoeijmakers, H. J. et al. Atomic iron and titanium in the atmosphere of the exoplanet KELT-9b. *Nature* **560**, 453 (2018).
12. Hoeijmakers, H. J. et al. A spectral survey of an ultra-hot Jupiter. Detection of metals in the transmission spectrum of KELT-9 b. *Astron. Astrophys.* **627**, A165 (2019).
13. Casasayas-Barris, N. et al. Atmospheric characterization of the ultra-hot Jupiter MASCARA-2b/KELT-20b. Detection of CaII, FeII, NaI, and the Balmer series of H (Hα, Hβ, and Hγ) with high-dispersion transit spectroscopy. *Astron. Astrophys.* **628**, A9 (2019).
14. Arcangeli, J. et al. Climate of an ultra hot Jupiter. Spectroscopic phase curve of WASP-18b with HST/WFC3. *Astron. Astrophys.* **625**, A136 (2019).
15. Zhang, J., Kempton, E. M. R. & Rauscher, E. Constraining hot Jupiter atmospheric structure and dynamics through Doppler-shifted emission spectra. *Astrophys. J.* **851**, 84 (2017).
16. Snellen, I. A. G., de Kok, R. J., de Mooij, E. J. W. & Albrecht, S. The orbital motion, absolute mass and high-altitude winds of exoplanet HD209458b. *Nature* **465**, 1049–1051 (2010).
17. West, R. G. et al. Three irradiated and bloated hot Jupiters: WASP-76b, WASP-82b, and WASP-90b. *Astron. Astrophys.* **585**, A126 (2016).
18. Brown, D. J. A. et al. Rossiter-McLaughlin models and their effect on estimates of stellar rotation, illustrated using six WASP systems. *Mon. Not. R. Astron. Soc.* **464**, 810–839 (2017).
19. Seidel, J. V. et al. Hot exoplanet atmospheres resolved with transit spectroscopy (HEARTS). II. A broadened sodium feature on the ultra-hot giant WASP-76b. *Astron. Astrophys.* **623**, A166 (2019).
20. Pepe, F. et al. ESPRESSO: the next European exoplanet hunter. *Astron. Nachr.* **335**, 8 (2014).
21. Queloz, D. et al. Detection of a spectroscopic transit by the planet orbiting the star HD209458. *Astron. Astrophys.* **359**, L13 (2000).
22. Cegla, H. M. et al. The Rossiter-McLaughlin effect reloaded: probing the 3D spin-orbit geometry, differential stellar rotation, and the spatially-resolved stellar spectrum of star-planet systems. *Astron. Astrophys.* **588**, A127 (2016).
23. Borsa, F. et al. The GAPS programme with HARPS-N at TNG XIX. Atmospheric Rossiter-McLaughlin effect and improved parameters of KELT-9b. *Astron. Astrophys.* **631**, A34 (2019).
24. Di Gloria, E., Snellen, I. A. G. & Albrecht, S. Using the chromatic Rossiter-McLaughlin effect to probe the broadband signature in the optical transmission spectrum of HD 189733b. *Astron. Astrophys.* **580**, A84 (2015).
25. Garhart, E. et al. Statistical characterization of hot Jupiter atmospheres using Spitzer's secondary eclipses. *Astron. J.* **159**, 137 (2020).
26. Showman, A. P. et al. Atmospheric circulation of hot Jupiters: coupled radiative-dynamical general circulation model simulations of HD 189733b and HD 209458b. *Astrophys. J.* **699**, 564–584 (2009).
27. Rauscher, E. & Menou, K. Three-dimensional modeling of hot Jupiter atmospheric flows. *Astrophys. J.* **714**, 1334–1342 (2010).
28. Beatty, T. G. et al. Spitzer phase curves of KELT-1b and the signatures of nightside clouds in thermal phase observations. *Astron. J.* **158**, 166 (2019).

29. Keating, D., Cowan, N. B. & Dang, L. Uniformly hot nightside temperatures on short-period gas giants. *Nat. Astron.* **3**, 1092–1098 (2019).
30. Malik, M. et al. Self-luminous and irradiated exoplanetary atmospheres explored with HELIOS. *Astron. J.* **157**, 170 (2019).
31. Santos, N. C., Israelian, G. & Mayor, M. Spectroscopic [Fe/H] for 98 extra-solar planet-host stars. Exploring the probability of planet formation. *Astron. Astrophys.* **415**, 1153–1166 (2004).
32. Sousa, S. G. et al. Spectroscopic parameters for 451 stars in the HARPS GTO planet search program. Stellar [Fe/H] and the frequency of exo-Neptunes. *Astron. Astrophys.* **487**, 373–381 (2008).
33. Sousa, S. G. et al. SWEET-Cat updated. New homogenous spectroscopic parameters. *Astron. Astrophys.* **620**, A58 (2018).
34. Sousa, S. G. in *Determination of Atmospheric Parameters of B-, A-, F- and G-Type Stars* (eds Niemczura, E., Smalley, B. & Pych, W.) 297–310 (GeoPlanet: Earth and Planetary Sciences, Springer, 2014).
35. Da Silva, R. et al. Elodie metallicity-biased search for transiting hot Jupiters. I. Two hot Jupiters orbiting the slightly evolved stars HD 118203 and HD 149143. *Astron. Astrophys.* **446**, 717–722 (2006).
36. Bressan, A. et al. PARSEC: stellar tracks and isochrones with the PAdova and TRieste Stellar Evolution Code. *Mon. Not. R. Astron. Soc.* **427**, 127–145 (2012).
37. Ginski, C. et al. A lucky imaging multiplicity study of exoplanet host stars - II. *Mon. Not. R. Astron. Soc.* **457**, 2173–2191 (2016).
38. Woellert, M. & Brandner, W. A lucky imaging search for stellar sources near 74 transit hosts. *Astron. Astrophys.* **579**, A129 (2015).
39. Lendl, M. et al. WASP-42 b and WASP-49 b: two new transiting sub-Jupiters. *Astron. Astrophys.* **544**, A72 (2012).
40. Lendl, M. et al. Signs of strong Na and K absorption in the transmission spectrum of WASP-103b. *Astron. Astrophys.* **606**, A18 (2017).
41. Espinoza, N. & Jordán, A. Limb darkening and exoplanets: testing stellar model atmospheres and identifying biases in transit parameters. *Mon. Not. R. Astron. Soc.* **450**, 1879–1899 (2015).
42. Mandel, K. & Agol, E. Analytic light curves for planetary transit searches. *Astrophys. J.* **580**, L171 (2002).
43. Haario, H., Saksman, E. & Tamminen, J. An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242 (2001).
44. Delisle, J. B. et al. The HARPS search for southern extra-solar planets. XLIII. A compact system of four super-Earth planets orbiting HD 215152. *Astron. Astrophys.* **614**, A133 (2018).
45. Foreman-Mackey, D., Agol, E., Ambikasaran, S. & Angus, R. Fast and scalable Gaussian process modeling with applications to astronomical time series. *Astron. J.* **154**, 220 (2017).
46. Sokal, A. D. in *Functional Integration* (eds Dewitt-Morette, C. & Folacci, A.) 131–192 (NATO ASI Ser. **Vol. 361**, Springer, 1997).
47. Foreman-Mackey, D. corner.py: scatterplot matrices in Python. *J. Open Source Softw.* **1**, 24 (2016).
48. Hansen, B. M. S. Calibration of equilibrium tide theory for extrasolar planet systems. *Astrophys. J.* **723**, 285–299 (2010).
49. Leconte, J., Chabrier, G., Baraffe, I. & Levrard, B. Is tidal heating sufficient to explain bloated exoplanets? Consistent calculations accounting for finite initial eccentricity. *Astron. Astrophys.* **516**, A64 (2010).
50. Bolmont, E., Raymond, S. N. & Leconte, J. Tidal evolution of planets around brown dwarfs. *Astron. Astrophys.* **535**, A94 (2011).
51. Gallet, F., Bolmont, E., Mathis, S., Charbonnel, C. & Amard, L. Tidal dissipation in rotating low-mass stars and implications for the orbital evolution of close-in planets. I. From the PMS to the RGB at solar metallicity. *Astron. Astrophys.* **604**, A112 (2017).
52. Bourrier, V., Cegla, H. M., Lovis, C. & Wyttenbach, A. Refined architecture of the WASP-8 system: a cautionary tale for traditional Rossiter-McLaughlin analysis. *Astron. Astrophys.* **599**, A33 (2017).
53. Bourrier, V. et al. Orbital misalignment of the Neptune-mass exoplanet GJ 436b with the spin of its cool star. *Nature* **553**, 477–480 (2018).
54. Lavie, B. et al. HELIOS–RETRIEVAL: an open-source, nested sampling atmospheric retrieval code; application to the HR 8799 exoplanets and inferred constraints for planet formation. *Astron. J.* **154**, 91 (2017).
55. Winn, J. N., Fabrycky, D., Albrecht, S. & Johnson, J. A. Hot stars with hot Jupiters have high obliquities. *Astrophys. J.* **718**, L145–L149 (2010).
56. Suárez Mascareño, A., Rebolo, R., González Hernández, J. I. & Esposito, M. Rotation periods of late-type dwarf stars from time series high-resolution spectroscopy of chromospheric indicators. *Mon. Not. R. Astron. Soc.* **452**, 2745–2756 (2015).
57. Suárez Mascareño, A., Rebolo, R. & González Hernández, J. I. Magnetic cycles and rotation periods of late-type stars from photometric time series. *Astron. Astrophys.* **595**, A12 (2016).
58. Hebb, L. et al. WASP-12b: the hottest transiting extrasolar planet yet discovered. *Astrophys. J.* **693**, 1920–1928 (2009).
59. Southworth, J. et al. High-precision photometry by telescope defocusing - VII. The ultrashort period planet WASP-103. *Mon. Not. R. Astron. Soc.* **447**, 711–721 (2015).
60. Delrez, L. et al. WASP-121 b: a hot Jupiter close to tidal disruption transiting an active F star. *Mon. Not. R. Astron. Soc.* **458**, 4025–4043 (2016).
61. Talens, G. J. J. et al. MASCARA-2 b: a hot Jupiter transiting the $m_V = 7.6$ A-star HD 185603. *Astron. Astrophys.* **612**, A57 (2018).
62. Kreidberg, L. et al. Global climate and atmospheric composition of the ultra-hot Jupiter WASP-103b from HST and Spitzer phase curve observations. *Astron. J.* **156**, 17 (2018).
63. Arcangeli, J. et al. H^- opacity and water dissociation in the dayside atmosphere of the very hot gas giant WASP-18b. *Astrophys. J.* **855**, L30 (2018).
64. Dobbs-Dixon, I., Cumming, A. & Lin, D. N. C. Radiative hydrodynamic simulations of HD209458b: temporal variability. *Astrophys. J.* **710**, 1395 (2010).
65. Heng, K., Frierson, D. M. W. & Phillipps, P. J. Atmospheric circulation of tidally locked exoplanets: II. Dual-band radiative transfer and convective adjustment. *Mon. Not. R. Astron. Soc.* **418**, 2669–2696 (2011).
66. Louden, T. & Wheatley, P. J. Spatially resolved eastward winds and rotation of HD 189733b. *Astrophys. J.* **814**, L24 (2015).
67. Helling, C. et al. Understanding the atmospheric properties and chemical composition of the ultra-hot Jupiter HAT-P-7b: I. Cloud and chemistry mapping. Preprint at http://arXiv.org/abs/1906.08127 (2019).

[1]Observatoire Astronomique de l'Université de Genève, Versoix, Switzerland. [2]Centro de Astrobiología (CSIC-INTA), Torrejón de Ardoz, Spain. [3]INAF Osservatorio Astronomico di Trieste, Trieste, Italy. [4]Instituto de Astrofísica de Canarias, Tenerife, Spain. [5]Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Porto, Portugal. [6]Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal. [7]INAF Osservatorio Astronomico di Brera, Merate, Italy. [8]Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, Lisbon, Portugal. [9]Departamento de Física da Faculdade de Ciências da Univeridade de Lisboa, Lisbon, Portugal. [10]Physikalisches Institut and Center for Space and Habitability, Universität Bern, Bern, Switzerland. [11]European Southern Observatory, Garching bei München, Germany. [12]Scuola Normale Superiore, Pisa, Italy. [13]European Southern Observatory, Santiago de Chile, Chile. [14]Space Research Institute, Austrian Academy of Sciences, Graz, Austria. [15]Centro de Astrofísica da Universidade do Porto, Porto, Portugal. [16]INAF Osservatorio Astronomico di Palermo, Palermo, Italy. [17]Institute for Fundamental Physics of the Universe, Miramare, Italy. [18]Fundación Galileo Galilei, INAF, Breña Baja, Spain. [19]INAF Osservatorio Astrofisico di Torino, Pino Torinese, Italy. ✉e-mail: david.ehrenreich@unige.ch

Methods

Observation log

The observations were carried out as part of the ESPRESSO Guaranteed Time Observation programme 1102.C-744. The observing conditions for both epochs (seeing, air masses, signal-to-noise ratios versus time) are reported in Extended Data Fig. 1. The seeing at the beginning of the observations was better in epoch 2 (about 0.9 arcsec) than in epoch 1 (about 1.3 arcsec). This explains why we opted for a longer exposure time in epoch 1. We excluded exposures obtained at an air mass above 2.2 as the atmospheric dispersion corrector (ADC) cannot handle correctly higher air masses.

Stellar parameters

WASP-76 (RA 01 h 46 min, dec. +02° 42′) is a F7 star of magnitude $V = 9.5$. Its properties were studied in the discovery paper[17], which used photometry from SuperWASP-North, WASP-South, TRAPPIST and EulerCAM at the Euler telescope, and spectroscopy from CORALIE at the Euler telescope and SOPHIE at the OHP 1.93-m telescope. One spectroscopic transit was later observed with HARPS at the ESO 3.6-m telescope on 11 November 2012[18]. Ref. [19] recently reported two new HARPS spectroscopic transits (24 October 2017 and 22 November 2017), accompanied by simultaneous EulerCAM photometry. In the meantime, the star was observed by Gaia: ref. [19] used the Gaia DR2 values (parallax, magnitude and effective temperature) and the new EulerCAM photometry to re-assess the stellar parameters. This resulted in increased values for the stellar radius (hence, the radius of the planet) and the stellar mass compared to the ones previously reported[17,18], which are based on a combined analysis of the photometric and spectroscopic data.

We performed a new analysis of the stellar parameters based on the ESPRESSO spectra and the Gaia DR2 parallax ($\pi = 5.12 \pm 0.15$ mas). For this, we combined several of our spectra to obtain a high signal-to-noise ratio spectrum of ~1,200 per resolution element, which we analysed using ARES+MOOG following a well-established spectroscopic analysis method[31–34]. The new stellar parameters we derive are listed in Extended Data Table 1. In particular, we obtain a stellar effective temperature of 6,329 ± 65 K that is compatible with ref. [18] (6,250 ± 100 K), and a log(g) of 4.196 ± 0.106 dex. Using the Padova stellar model isochrones (http://stev.oapd.inaf.it/cgi-bin/param_1.3)[35,36] and the Gaia parallax, we obtain a stellar age of 1.816 ± 0.274 Gyr, a $B - V$ colour of 0.569 ± 0.017 mag, a stellar mass of $(1.458 \pm 0.021) M_{\odot}$ and a stellar radius of $(1.756 \pm 0.071) R_{\odot}$. The two latter values, strongly constrained by the Gaia parallax, are consistent with the values reported by Brown et al.[18] and we adopted them as the new default stellar parameters in the rest of this study.

Binary companion

WASP-76 has a candidate companion separated by 0.4438 ± 0.0053 arcsec and a magnitude difference of 2.58 ± 0.27 (refs. [37,38]), which corresponds to a flux contrast of ~10. By combining these ground-based measurements with Keck and HST/STIS images, it is possible to establish that the candidate companion is actually bound to the star (G. Fu, personal communication) and determine its effective temperature (~5,100 K) and radius ($0.8 R_{\odot}$). WASP-76B thus resembles a late G- or early K-type dwarf. This candidate companion lies at the limit of the entrance of the 0.5-arcsec-radius ESPRESSO fibre. Given the seeing values reported in Extended Data Fig. 1a, d, it most certainly contaminates our spectra and the contamination could vary with the seeing. We checked that the companion does not impact the stellar parameters by repeating our analysis selecting stellar spectra only in conditions of good seeing (≤0.85″) or bad seeing (>1″). The results are compatible with the values derived above within their stated uncertainties. We performed an extensive search for a spectroscopic signature of WASP-76B; however, we could not find any sign of a contamination of the stellar CCFs of WASP-76A down to the ~500 ppm level. Given its above-stated properties, it is surprising that the CCF of WASP-76B remains undetectable. One possibility is that WASP-76B is a fast rotator, producing a broad CCF that would be lost in the noise; however, the system is not particularly young. Another possibility is that the CCF of WASP-76B has almost identical radial velocity, FWHM and contrast as the CCF of the primary star, which would thus efficiently 'hide' it. Because it is so well hidden, WASP-76B is unlikely to affect our results (see also next section).

Transit photometry

We performed a new photometric analysis based on all six existing transit light curves of WASP-76b obtained with the EulerCam instrument at the Swiss Euler 1.2-m telescope in La Silla, Chile. We extracted the raw light curves (Extended Data Fig. 8a) using aperture photometry described in ref. [39]. We jointly analysed all photometric data sets using a differential-evolution Markov Chain Monte Carlo (MCMC) code[40], fitting for the mid-transit time, revolution period, planet-to-star radius ratio and system scale with Gaussian priors. We accounted for instrumental systematics and red noise (see ref. [19] for more details). We used a quadratic limb-darkening law and obtained the corresponding u_1 and u_2 coefficients for WASP-76 with the routines of ref. [41]. The corrected light curves are shown in Extended Data Fig. 8b and derived parameters are reported in Extended Data Table 1. These values take into account the dilution caused by WASP-76B. Its main effect is to increase slightly the transit depth and planet-to-star radius ratio. However, the effect of the planet radius is small; we found $1.863^{+0.070}_{-0.083}$ $R_{♃}$, where $R_{♃}$ is the radius of Jupiter. This value is actually smaller than the one previously reported in ref. [19]; this is because we made use of a smaller (and more accurate) stellar radius (see Methods section 'Stellar parameters'). We used our newly derived parameters to create a transit model[42] that we used to perform the reloaded Rossiter–McLaughlin analysis (see Methods section 'Rossiter effect and Doppler shadow') and to convert the residual amplitudes into differential transit depths (Extended Data Fig. 6c).

Cross-correlation mask

We used the built-in cross-correlation mask corresponding to an F9-type star in the ESPRESSO pipeline to obtain the stellar CCF for each exposure. This mask was created by collecting the position of all lines for the F9.5 star HD1581. The lines were individually identified by querying their wavelengths in the Vienna Atomic Line Database (VALD; http://vald.astro.uu.se). Iron (Fe) is by far the most represented element (47% of the 4,653 spectral lines). The next most-represented atoms are nickel (Ni, 6.5%), chromium (Cr, 5.7%) and titanium (Ti, 4.8%). The most represented ion in the F9 mask is Ti^{+} (2.8% of the lines). The CCF is then built as a flux-weighted and contrast-weighted mean line profile. Many deep (contrasted) Fe lines are located in spectral regions where the stellar (continuum) flux is high, therefore boosting the importance of the Fe lines. Their actual weight in the F9 mask represents 76% of the spectroscopic information.

Orbital solution

We retrieved the orbital parameters and planet mass from our ESPRESSO measurements. We excluded data points obtained during transits from the analysis to prevent the Rossiter–McLaughlin effect perturbing the orbital solution. This left most of our in-transit data aside, hence we collected new data points to extend our coverage of the orbit (in particular, at the quadratures) at high precision, during autumn 2019 (hereafter, epoch 3). The new radial velocities are presented in Extended Data Table 2 and Extended Data Fig. 2.

To derive the uncertainties on the orbital parameters from the ESPRESSO data, we computed their posterior distribution with a MCMC algorithm. We modelled the signal with a Keplerian and three radial velocity offsets: offsets 1 and 2 correspond to epochs 1 and 2, respectively. A technical intervention on ESPRESSO occurred between these two epochs and could have changed the reference 'zero' radial

velocity of the spectrograph; hence the need to introduce an offset. Offset 3 corresponds to the set of subsequent observations (spanning BJD 2,458,684 to 2,458,754) obtained to increase the precision of the radial velocity solution. Our variables are these three offsets, the period P, the semi-amplitude K_*, the eccentricity e, the argument of periastron ω and the inferior conjunction time T_{conj}. We parametrized the noise in the covariance matrix by three terms: two standard deviations for white (σ_W) and red (σ_R) noise and a characteristic timescale τ of the red noise. The prior distributions on all parameters are listed in Extended Data Table 3. The constraints on the period P and $e\cos\omega$ are obtained from the Spitzer[25] and Euler transits (see Methods section 'Transit photometry'). The MCMC algorithm is an adaptive Metropolis algorithm[43] as implemented in ref. [44]. We used a homegrown package (spleaf; J.-B. Delisle et al., submitted) based on an efficient numerical method[45] to speed up the covariance calculations. To check the convergence of the chain, we computed the number of effective samples from the autocorrelation function of the chain[44,46]. We obtained the posterior distribution of the planet mass by computing it as a function of P, K_*, e, the inclination i_* and the stellar mass M_*. We used the MCMC samples of P, K_* and e and sampled independently i_* and M_* from their constraints as given in Extended Data Table 1. The distribution of i_* is approximated by a mixture of two Gaussians (with mean $\mu = 89.623°$, and standard deviations $\sigma_1 = 0.034°$ and $\sigma_2 = 0.005°$). The distribution of M_* is approximated by a Gaussian distribution ($\mu = 1.458 M_\odot$, $\sigma = 0.021 M_\odot$). The maximum-likelihood fit, the posterior median and the 1σ confidence intervals are given in Extended Data Table 3. The corner plot, made with the corner.py code[47], is presented in Extended Data Fig. 3 and the radial velocity data with the maximum-likelihood fit in Extended Data Fig. 2.

The maximum likelihood of the eccentricity and its posterior median are close to 0. Previous studies based on CORALIE data have adopted a null eccentricity[17,18] and our more precise ESPRESSO measurements also point towards a circular orbit. The best constraint comes from the Spitzer measurement[25] of $e\cos\omega = -0.00135 \pm 0.00083$, which also gives a strong indication of a null eccentricity for the most likely (small) values of ω. A null eccentricity is also strongly favoured by theory, considering the expected short circularization timescale: an equilibrium tide[48] would damp an eccentricity of 10% in about 30 Myr. Given the age of the star (1.8 ± 0.3 Gyr), there would have been ample time for the orbit to fully circularize, especially considering that a dynamical tide in the fluid layers of the planet would result in a higher dissipation factor. Note that the stellar tide could potentially excite the eccentricity if its rotation is fast enough (if the spin of the star is larger than 18/11 of the orbital frequency[49,50]). This is not the case here since the stellar rotation frequency (0.03 d^{-1}; see Methods section 'Spin–orbit angle and stellar rotation') is much smaller than $18/11 \times P^{-1} = 0.9 d^{-1}$. This was also probably the case in the past for much longer than 30 Myr (see figure 2 in ref. [51]). Considering all of this, we decided to fix the eccentricity to 0.

Rossiter effect and Doppler shadow

The idea of the reloaded Rossiter–McLaughlin effect is to directly track the stellar surface radial velocity behind the transiting planet[22,52,53]. Following ref. [22], we shifted the stellar CCFs into the stellar rest frame. For this, we made use of the orbital solution obtained above. Since ESPRESSO observations are not flux-calibrated, the continuum levels of the CCFs are arbitrary. We normalized each CCF by its continuum level determined from the Gaussian fit. We then scaled each normalized CCF according to its timing with respect to the planetary transit. For this, we calculated the theoretical transit light curve as described in Methods section 'Transit photometry'. We shifted (in velocity) the rescaled CCFs by the velocity measured for a mean 'master' out-of-transit CCF. This is done so that the final result is independent of the velocity offsets (systemic velocity as well as instrument offsets are discussed in Methods section 'Orbital solution'). Finally, we produced the CCFs of the occulted stellar surface ('local' CCFs) by subtracting each scaled in-transit CCF from the master out-of-transit CCF.

The projected velocities of the stellar surface behind the planet during the transit appear first blueshifted, then redshifted, which indicates a prograde planetary orbit. The surface velocities roughly follow a straight line, indicative of solid-body rotation. We verified this using a dedicated stellar rotation model described below, which we adjusted to the data, with the exception of the flattened portion seen after mid-transit, at a time when the planet shimmer intersects with the Doppler shadow. We also excluded local CCFs where the stellar line was not detected at more than 5σ; these CCFs have been obtained at the ingress and egress.

Spin–orbit angle and stellar rotation

We fitted the stellar surface velocities with a model of stellar surface rotation assuming solid-body rotation[22]. For WASP-76, there is a known degeneracy between the projected spin–orbit angle λ, the projected equatorial rotational velocity of the star $v_{eq}\sin i_*$ (where i_* is the inclination of the stellar spin with respect to the plane of the sky) and the impact parameter[18] due to the very small value of the impact parameter (the transit is almost central). The impact parameter can be expressed as $a/R_*\cos i_p$, where i_p is the inclination of the planetary orbit. We chose λ, $v_{eq}\sin i_*$, a/R_* and i_p as free parameters. We embedded the model in a nested sampling retrieval algorithm[54] to efficiently explore the full parameter space. The priors on the four parameters were set as: (i) a uniform prior on λ ranging from −180° to 180°, (ii) a Gaussian prior ($\mu = 1.61$ km s^{-1}, $\sigma = 0.28$ km s^{-1}) on $v_{eq}\sin i_*$, which we derived as the quadratic difference between the FWHM of the stellar local master CCF and the FWHM of the stellar master-out CCF; (iii, iv) the posterior distributions of the Euler photometry (see Methods section 'Transit photometry') were chosen as priors for a/R_* and i_p. We performed a run of 5,000 living points; the best-fit parameters are the ones maximizing the logarithm of the evidence, $\log \mathcal{Z}$. The maximum $\log \mathcal{Z}$ of 8.59 ± 0.04 was obtained for $\lambda = 61.28°^{+7.61°}_{-5.06°}$, $v_{eq}\sin i_* = 1.48 \pm 0.28$ km s^{-1}, $a/R_* = 4.09 \pm 0.07$ and $i_p = 89.74°^{+0.15°}_{-0.11°}$. The quoted 1σ uncertainties are obtained from the posterior distributions of the parameters, which are shown in Extended Data Fig. 4. Based on the value of λ and the host star effective temperature of 6,329 K, WASP-76b lies at the transition between aligned and misaligned hot gas giants[55]. The slow (projected) rotation velocity we derived hints at a non-negligible inclination i_* of the stellar spin axis towards the line of sight (which is not constrained by the Rossiter–McLaughlin effect). In fact, $i_* = 90°$ would yield a rotation period of 60 days, which is much larger than the typical range of about 15–40 days expected for F stars[56,57]. A light curve of WASP-76 obtained with the All Sky Automated Survey for Supernovæ (ASAS) hints at a periodicity of about 35 days, which would yield an inclination of $i_* \approx 36°$ with respect to the line of sight, that is, the star would be close to pole-on.

Evening to dayside temperature rise

Ultrahot gas giants are an emerging class of exoplanets. In addition to WASP-76b, some of its representative objects are WASP-12b[58], WASP-33b[1], WASP-103b[59], WASP-121b[60], MASCARA-2b[61] and KELT-9b[2]. Observations have enabled the measurement of some of their physical and chemical properties, such as their temperature structures or composition[3,10–14,19,62,63]; however, no consistent picture of these extreme climates exists yet, as interpretations are essentially based on global circulation models established for less-irradiated hot gas giants[7,8,15,26,27,64,65]. While these models can be used to interpret wind measurements in planets like HD 209458b[16] or HD 189733b[66], they are less adapted to objects like WASP-76b. Recent theoretical developments aimed at understanding ultrahot atmospheres[4–6,9,67] and future work will be the basis on which to finely interpret spectroscopically and temporally resolved measurements such as the ones presented here.

In particular, we exploit here the idea that atoms recombine into molecules across the evening terminator and that molecules dissociate into atoms across the morning terminator of an ultrahot gas giant[9].

Applying this to iron atoms make it possible to use the Fe signature as a thermometer. The lowest absorption depth of $\delta_{\rm atm}^{\rm eve} = 434 \pm 32$ ppm is measured when only the evening terminator contributes to the signal (between steps 1 and 2 of the scenario depicted in Fig. 3). This absorption depth at the evening terminator can be expressed as $\delta_{\rm atm}^{\rm eve} \approx 2(R_{\rm p}/R_*)^2(H^{\rm eve}/R_{\rm p})n_{\rm H}$, where $H^{\rm eve}$ is the atmospheric scale height at the evening terminator and $n_{\rm H}$ is the number of scale heights over which the absorption takes place. The largest absorption depth of $\delta_{\rm atm}^{\rm day} = 628 \pm 49$ ppm is observed at the end of the transit, when the line of sight through the trailing limb probes regions close to the dayside hotspot (step 3 in Fig. 3). Assuming the absorption signal takes place over the same number of scale heights as on the evening terminator, we can write $\delta_{\rm atm}^{\rm day}/\delta_{\rm atm}^{\rm eve} = H^{\rm day}/H^{\rm eve}$, where $\delta_{\rm atm}^{\rm day}$ and $H^{\rm day}$ are the absorption depth and atmospheric scale height on the planet dayside, respectively. Since $H = kT/\mu g$, where k is Boltzmann's constant, μ is the mean molecular mass of the atmosphere, g is the surface gravity and T is the temperature, we obtain $\delta_{\rm atm}^{\rm day}/\delta_{\rm atm}^{\rm eve} = T^{\rm day}/T^{\rm eve} \approx 1.5 \pm 0.2$, where $T^{\rm day}$ and $T^{\rm eve}$ are the temperatures on the dayside and evening terminator, respectively. An occultation of WASP-76b by its host star was observed with Spitzer, providing a measurement of the dayside brightness temperature at 3.6 μm: 2,693 ± 56 K (ref. [25]). Using this value and the variation of the absorption depth measured here, we can estimate that WASP-76b has an evening terminator temperature of 1,795 ± 242 K.

Data availability

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available owing to the proprietary status of data obtained in the framework of the ESPRESSO Guaranteed Time Observations. At the end of the proprietary period, the data will be publicly available in the ESO archive (https://archive.eso.org).

Code availability

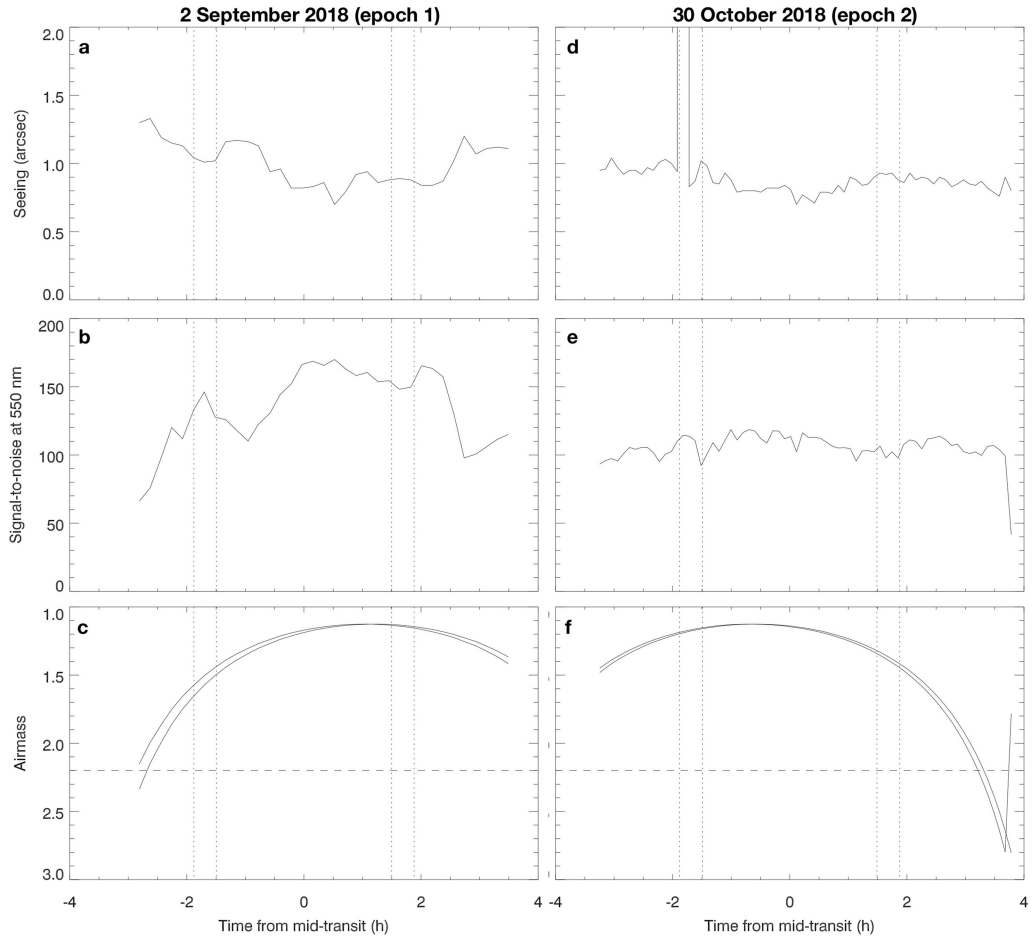
The ESPRESSO DRS is public software available from ESO at https://www.eso.org/sci/software/pipelines/espresso/espresso-pipe-recipes.html. The main analysis routines have been written by the authors in Interactive Data Language and are available upon reasonable request from the corresponding author.

Acknowledgements We thank G. Fu for sharing information regarding the binary companion of WASP-76A and D. Kitzmann for discussion about how iron can condense. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project FOUR ACES; grant agreement no. 724427). It has also been carried out in the frame of the National Centre for Competence in Research PlanetS supported by the Swiss National Science Foundation (SNSF). This work was supported by FCT/MCTES through national funds and by FEDER (Fundo Europeu de Desenvolvimento Regional) through COMPETE2020 (Programa Operacional Competitividade e Internacionalização) by these grants: UID/FIS/04434/2019; PTDC/FIS-AST/32113/2017 and POCI-01-0145-FEDER-032113; PTDC/FIS-AST/28953/2017 and POCI-01-0145-FEDER-028953. V.A. and S.S. acknowledge support from FCT through Investigador FCT contracts IF/00650/2015 and IF/00028/2014, and POPH/FSE (EC) by FEDER funding through the programme "Programa Operacional de Factores de Competitividade—COMPETE". This work of C.J.A.P.M. was financed by FEDER funds through the COMPETE 2020—Operational Programme for Competitiveness and Internationalisation (POCI), and by Portuguese funds through FCT (Fundação para a Ciência e a Tecnologia) in the framework of the projects POCI-01-0145-FEDER-028987 and UID/FIS/04434/2019. O.D. is supported by a work contract (DL 57/2016/CP1364/CT0004). M.R.Z.O. acknowledges financial support from AYA2016-79425-C3-2-P from the Spanish Ministry for Science, Innovation and Universities (MICIU). J.I.G.H. acknowledges financial support from the MICIU under the 2013 Ramón y Cajal programme MICIU RYC-2013-14875. J.I.G.H., R.R., C.A.P. and A. Suárez Mascareño also acknowledge financial support from the MICIU for project AYA2017-86389-P. This publication makes use of The Data and Analysis Center for Exoplanets (DACE), which is a facility based at the University of Geneva (CH) dedicated to extrasolar planet data visualization, exchange and analysis. DACE is a platform of the Swiss National Centre of Competence in Research (NCCR) PlanetS, federating the Swiss expertise in Exoplanet research. The DACE platform is available at https://dace.unige.ch. This paper is based on observations made at the ESO Very Large Telescope (Paranal, Chile) under programme 1102.C-744 and at the ESO 3.6-m telescope (La Silla, Chile) under programmes 090.C-0540 and 100.C-0750.

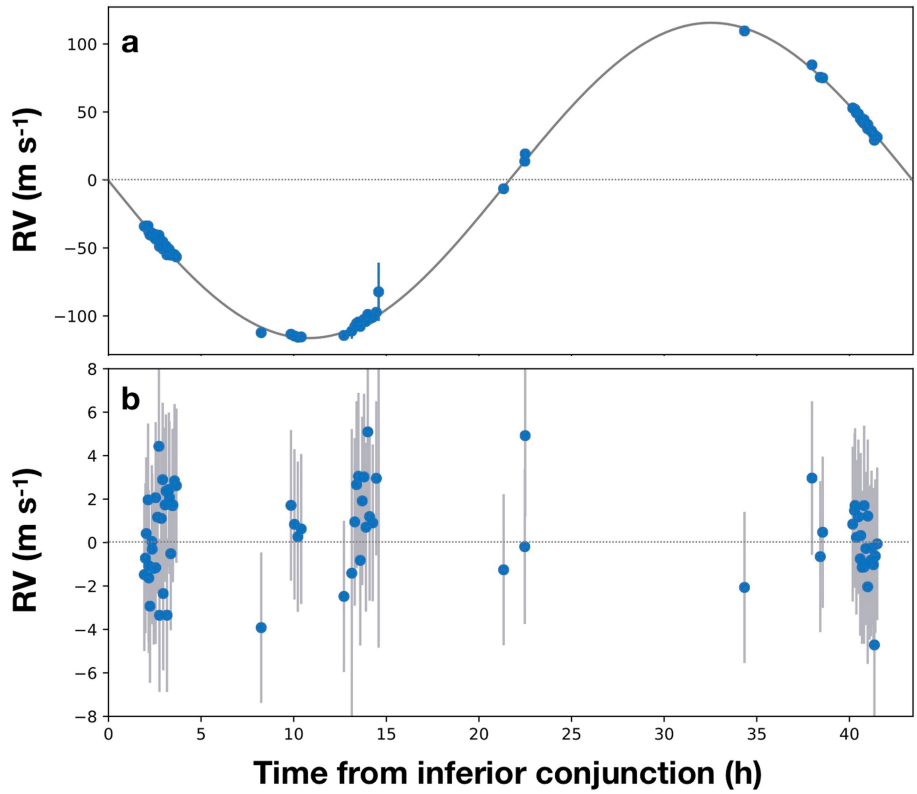
Author contributions D.E., C.L. and R.A. led the data analysis and interpretation. D.E. wrote the paper with contributions from R.A. C.L. led the development of the data reduction pipeline. M.R.Z.O. coordinated the observations and scientific work and performed the first-epoch observation. F.P., S.C., R.R. and N.C.S. led the ESPRESSO consortium and building of the instrument. J.I.G.H. performed the second-epoch observation. F. Borsa, O.D., E. Pallé, N.C.S., E.B., V. Bourrier, H.M.C., N.C.-B., J.V.S. and H.T. brought decisive contributions to the interpretation. N.C.-B. performed an independent data analysis. S.S. performed the stellar parameter analysis. X.D. created the CCF mask and retrieved the list of its atomic lines. N.H. made the radial velocity retrieval. D. Ségransan provided support with DACE. B.L. provided the nested sampling algorithm for the analysis. M. Lendl derived the transit ephemeris. V.A., C.A.P., Y.A., F. Bouchy, V.D., P.F., R.G.S., C.J.A.P.M., A. Mehner, G.M., P.M., N.N., G.L.C., E. Poretti, A.S., A. Suárez Mascareño and S.U. participated in the scientific preparation and target selection for these observations. The other co-authors provided key contributions to the instrumental, software and operational development of ESPRESSO. All co-authors read and commented the manuscript.

Competing interests The authors declare no competing interests.

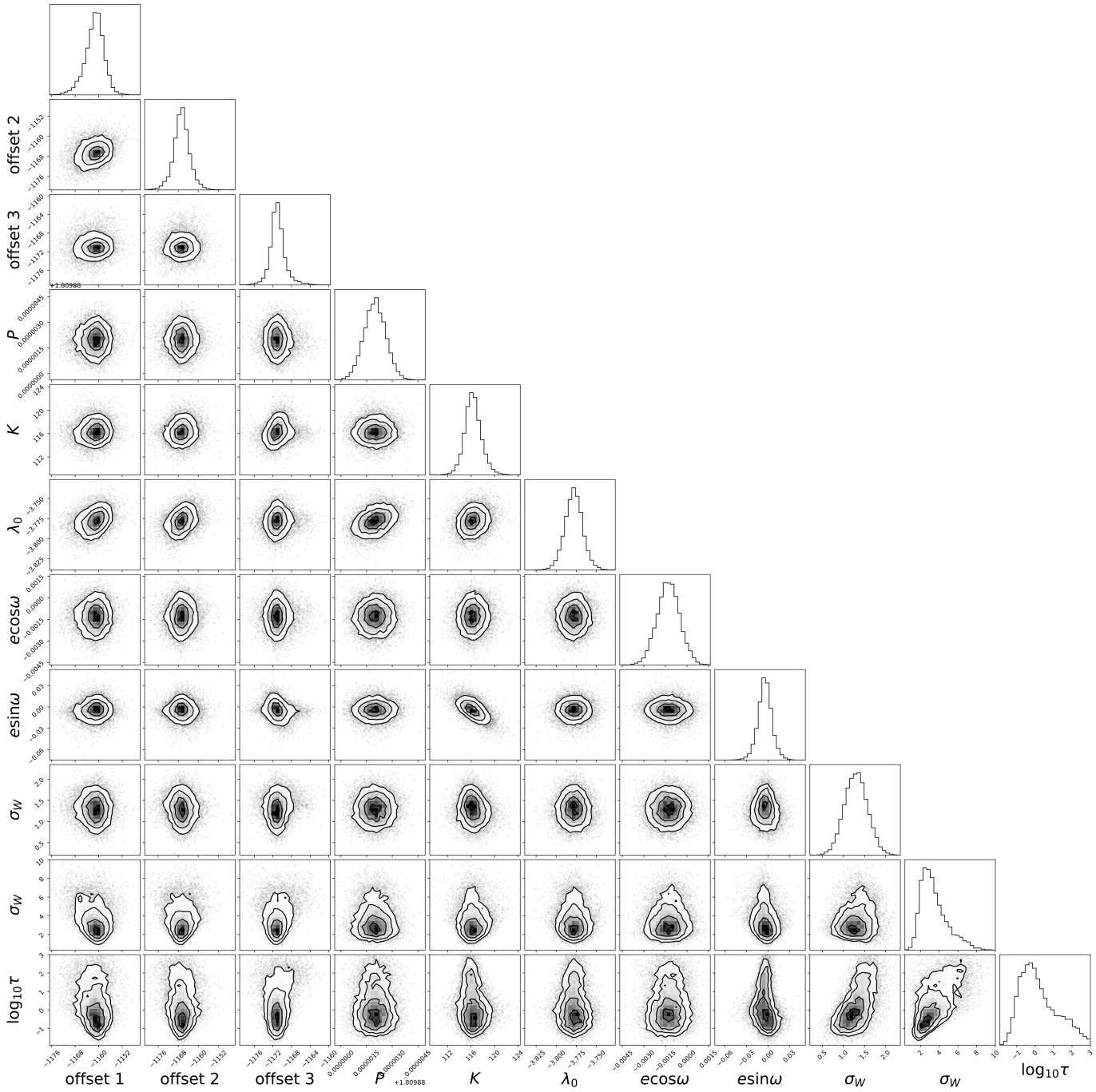
Additional information
Correspondence and requests for materials should be addressed to D.E.
Peer review information *Nature* thanks Drake Deming and Ignas Snellen for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



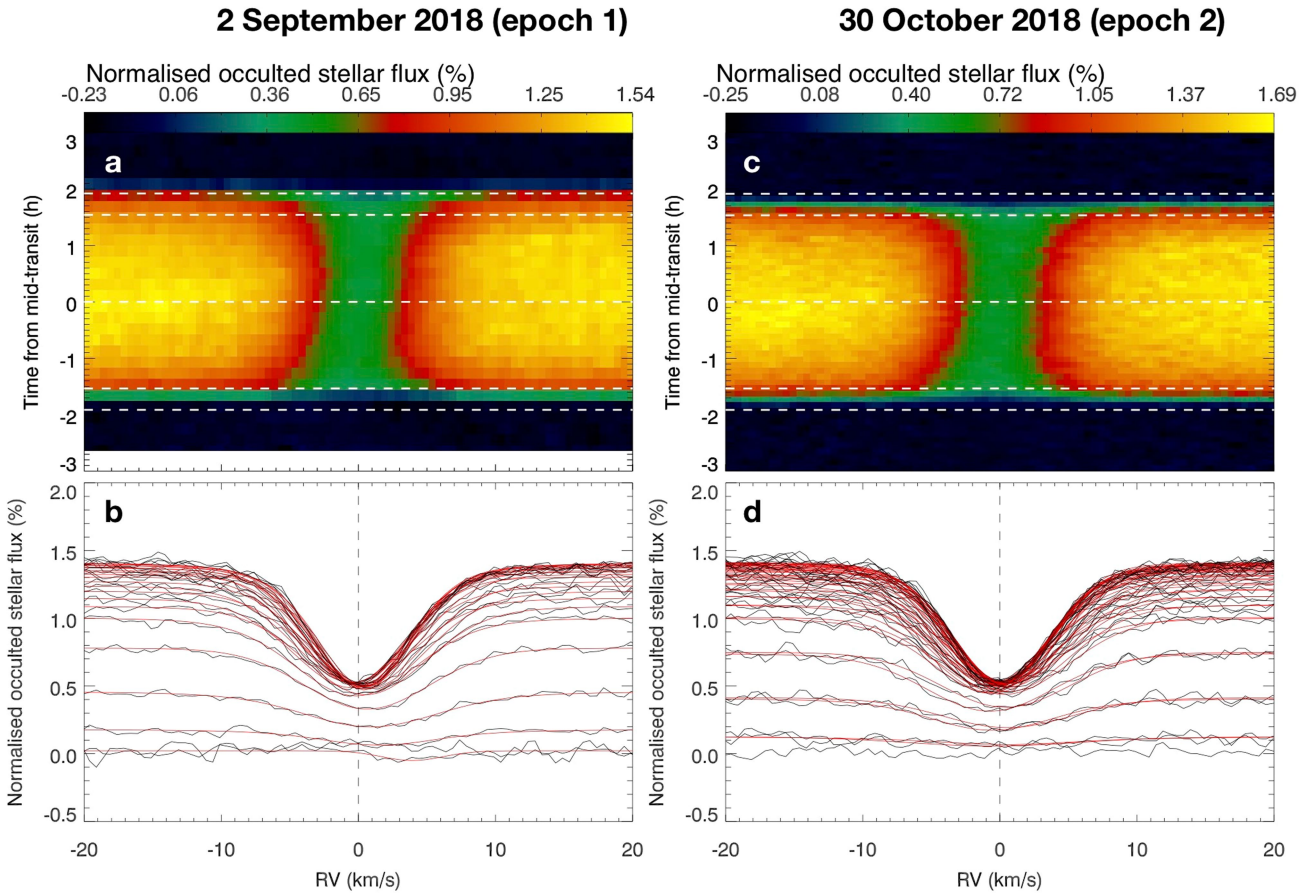
Extended Data Fig. 1 | Variations of the observing conditions during transit epochs 1 and 2. a–**c**, Epoch 1. **d**–**f**, Epoch 2. The seeing (**a**, **d**), signal-to-noise ratio per pixel at 550 nm (**b**, **e**) and airmass (**c**, **f**) are shown as a function of the time in transit. Vertical dotted lines represent the transit contacts. The horizontal dashed lines in **c** and **f** indicate the airmass of 2.2 beyond which the data are discarded from the analysis.



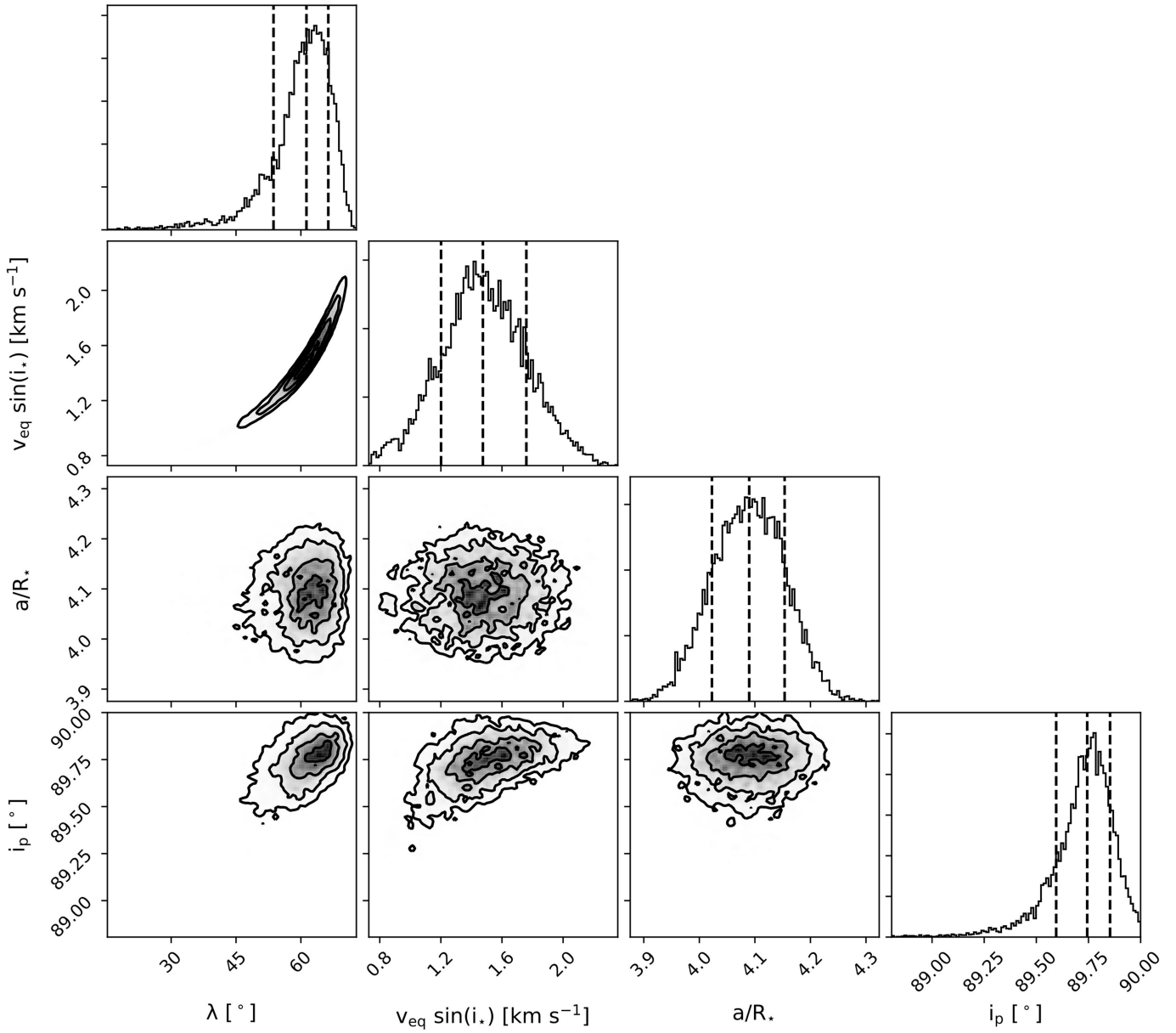
Extended Data Fig. 2 | ESPRESSO radial velocities of WASP-76. a, Stellar radial velocities (RV; blue points) and the maximum-likelihood fit using values from Extended Data Table 3. The transit occurs at the inferior conjunction (0 h). In-transit data have been removed as they are affected by the Rossiter–McLaughlin effect and the atmospheric absorption from the planet. **b**, Residuals of the radial velocities after subtraction of the maximum-likelihood fit. The standard deviation of the residuals is about 2.8 m s^{-1}. Error bars correspond to 1σ uncertainties and include both parameterized noise terms.



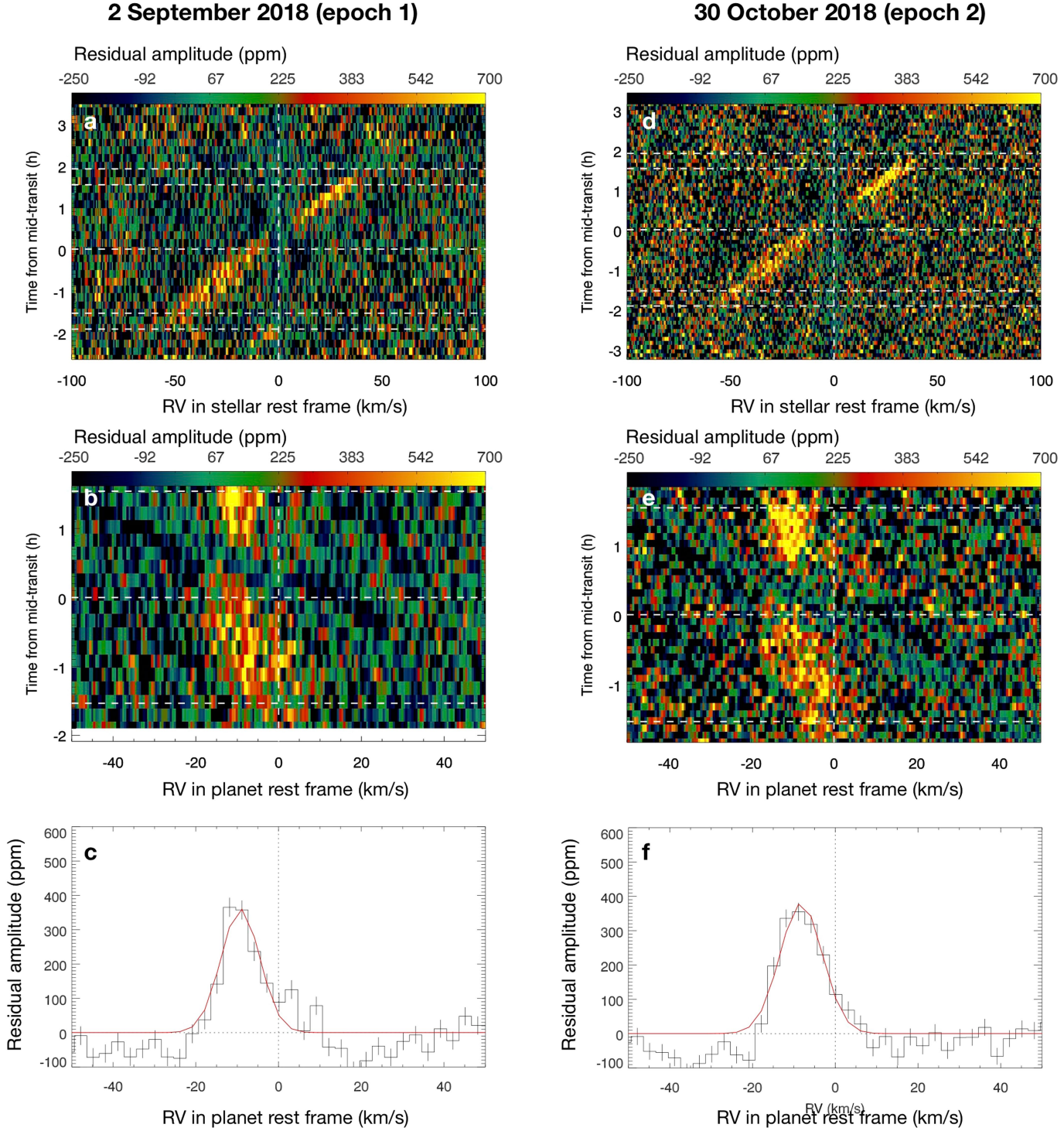
Extended Data Fig. 3 | MCMC chain corner plot. Shown is the corner plot for the orbital parameters representing the posterior distribution of variables used for the MCMC computations of the orbital parameters. The posterior distribution medians are reported in Extended Data Table 3.



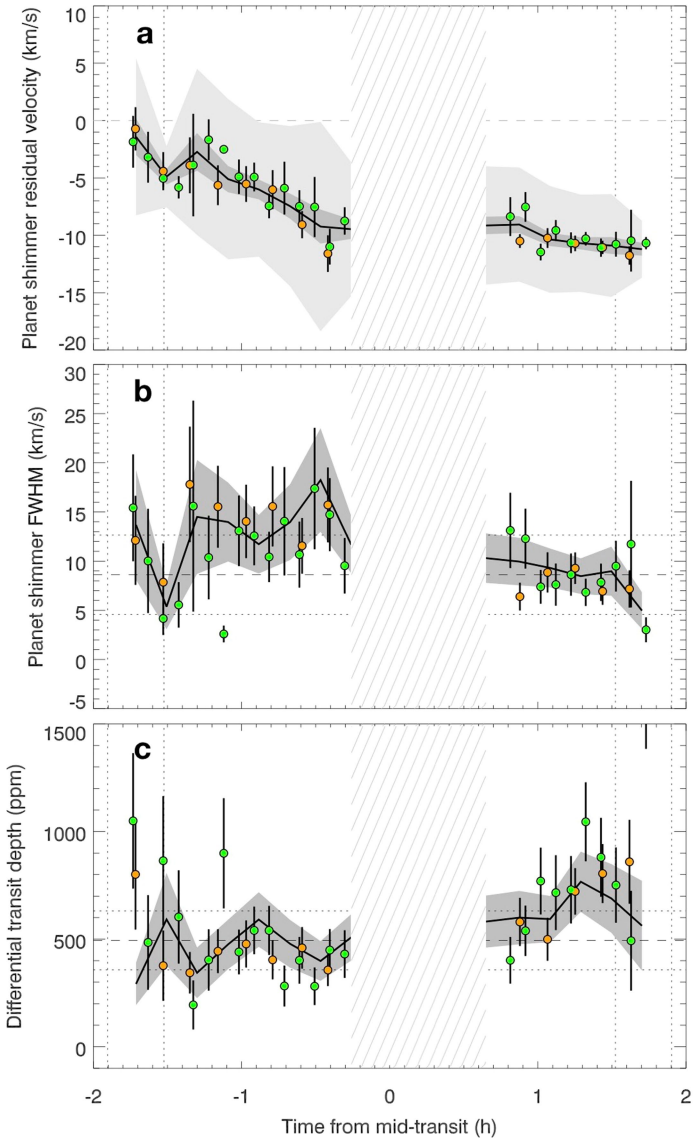
Extended Data Fig. 4 | Doppler shadow of WASP-76. a, **b**, Data for epoch 1; **c**, **d**, data for epoch 2. **a**, **c**, Local stellar CCFs behind the planet represented as a function of time. The horizontal dashed lines represent (from bottom to top) the second contact, mid-transit and third contact. **b**, **d**, 1D view of the local stellar CCFs (black lines) with their Gaussian fits (red curves).



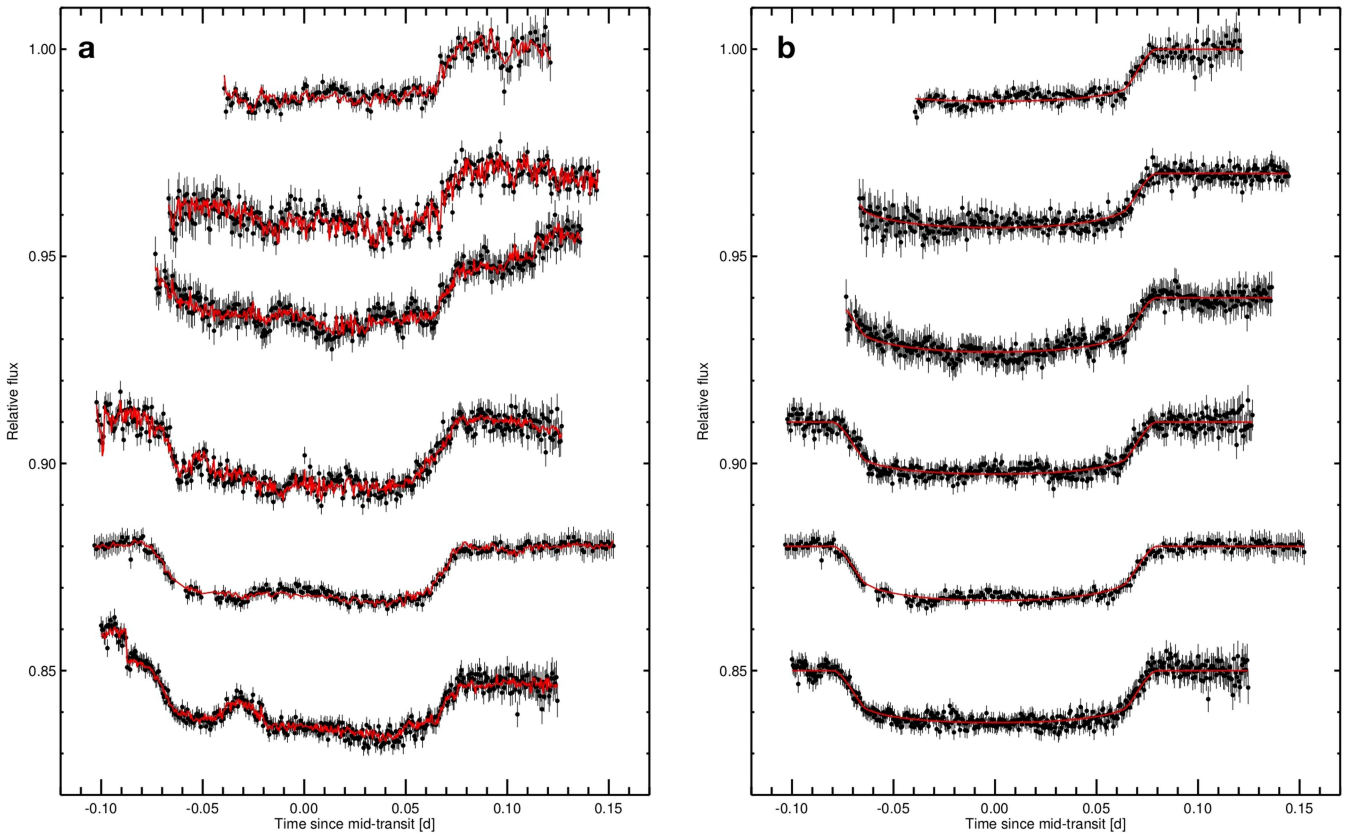
Extended Data Fig. 5 | Parameters of the stellar surface rotation model. The corner plot shows the posterior distributions of the four free parameters of the model, the projected spin–orbit angle λ, the projected equatorial stellar rotational velocity $v_{eq}\sin i_\star$, the system scale $a/R_\star$ and the planetary orbit inclination i_p. The posterior distribution medians and their 1σ uncertainties are represented by vertical dashed lines.



Extended Data Fig. 6 | Absorption signature of WASP-76b. a–**c**, On 2 September 2018 (epoch 1); **d**–**f**, on 30 October 2018 (epoch 2). The planetary absorption signal is shown in the stellar rest frame (**a**, **d**), the planet rest frame (**b**, **e**) and is time-averaged in the planet rest frame to produce the atmospheric absorption profile integrated over the whole limb (**c**, **f**). An indicative Gaussian fit (red curves) is overplotted on the absorption profiles. Both epochs show compatible results.



Extended Data Fig. 7 | Measured properties of the planetary absorption signature as a function of time. Data from epoch 1 (orange), epoch 2 (green) and both epochs combined (binned by 2; black curve with 1σ uncertainty in dark grey) are shown. They result from Gaussian fits to the planetary absorption signal in the residual maps of Fig. 2b and Extended Data Figs. 5b and e. A factor of $(R_p/R_*)^2/(1-\Delta F/F(t))$ was applied to the residual maps before the fit, where $\Delta F/F(t)$ is the model light curve used to extract the Doppler shadow. **a**, Radial velocity of the planetary signal (called the planet 'shimmer') in the planet rest frame. The light grey region shows the FWHM associated with each point. **b**, The FWHM of the signal. The weighted mean (horizontal dashed line) is 8.6 ± 0.7 km s^{-1}. Horizontal dotted lines indicate the standard deviation of the values. **c**, Amplitude of the shimmer representing the differential transit depth. The weighted mean is 494 ± 27 ppm. The hatched area in all panels represents the overlap between the Doppler shadow and the planetary signal; data between −0.2 h and +0.7 h from mid-transit are excluded from the analysis. Error bars in all panels correspond to 1σ uncertainties.



Extended Data Fig. 8 | Photometric transit light curve of WASP-76b obtained with the EulerCam instrument on the Swiss Euler 1.2-m telescope in La Silla, Chile. The last three transits (bottom rows) have been previously reported[19]. **a**, Raw light curves with their best-fit models including systematic effects. **b**, Normalized light curves. Error bars represent 1σ uncertainties.

Extended Data Table 1 | Parameters for WASP-76 and its planet

Parameter	Unit	Value	Reference/Methods
Gaia DR2 ID	–	2512326349403275520	CDS Simbad
Right ascension (J2000)	hms	$01^h46^m31.9^s$	CDS Simbad
Declination (J2000)	dms	+02°42'02.0"	CDS Simbad
V	mag	9.52±0.03	CDS Simbad
Spectral type	–	F7	CDS Simbad
Systemic velocity, γ_{sys}	km s^{-1}	–1.11±0.50	CDS Simbad
Parallax, π	mas	5.12±0.16	*Gaia* DR2
Stellar properties derived from ESPRESSO spectra			
Stellar mass, $M_\star$	$M_\odot$	1.458±0.021	"Stellar parameters"
Stellar radius, $R_\star$	$R_\odot$	1.756±0.071	"Stellar parameters"
Effective temperature, $T_{\rm eff}$	K	6,329±65	"Stellar parameters"
Stellar surface gravity, log g	cgs	4.196±0.106	"Stellar parameters"
Turbulent velocity, $v_{\rm turb}$	km s^{-1}	1.543±0.027	"Stellar parameters"
Colour, B–V	mag	0.569±0.017	"Stellar parameters"
Metallicity, [Fe/H]	–	0.366±0.053	"Stellar parameters"
Age	Gyr	1.816±0.274	"Stellar parameters"
Projected equatorial rotational velocity, $v_{\rm eq} \sin i_\star$	km s^{-1}	1.48±0.28	"Spin-orbit angle"
Spin-orbit projected angle, λ	deg	$61.28^{+7.61}_{-5.06}$	"Spin-orbit angle"
System properties retrieved from radial velocities			
Eccentricity, e	–	0 (fixed)	"Orbital solution"
Semi-amplitude of the stellar RVs, $K_\star$	m s^{-1}	$116.02^{+1.29}_{-1.35}$	"Orbital solution"
Planet mass, M_p	$M_{♃}$	$0.894^{+0.014}_{-0.013}$	"Orbital solution"
Systemic velocity for epoch 1, γ_{181}	m s^{-1}	$-1{,}162.00^{+2.86}_{-2.63}$	"Orbital solution"
Systemic velocity for epoch 2, γ_{182}	m s^{-1}	$-1{,}167.54^{+2.79}_{-2.73}$	"Orbital solution"
Systemic velocity for epoch 3, γ_{193}	m s^{-1}	$-1{,}171.11^{+1.28}_{-1.36}$	"Orbital solution"
System properties retrieved from photometry			
Period P	days	$1.80988198^{+0.00000064}_{-0.00000056}$	"Transit photometry"
Mid-transit time T_c	BJD	$58080.626165^{+0.000418}_{-0.000367}$	"Transit photometry"
Radius ratio $R_p/R_\star$	–	$0.10852^{+0.00096}_{-0.00072}$	"Transit photometry"
System scale $a/R_\star$	–	$4.08^{+0.02}_{-0.06}$	"Transit photometry"
Inclination i	deg	$89.623^{+0.005}_{-0.034}$	"Transit photometry"
Phases of contacts I and IV, $\varphi_{1\text{-}4}$	–	±0.043	"Transit photometry"
Phases of contacts II and III, $\varphi_{2\text{-}3}$	–	±0.034	"Transit photometry"
Ingress duration, ΔT_{12}	min	23.6	"Transit photometry"
Total transit duration, ΔT_{14}	min	230	"Transit photometry"
Transit depth	%	$1.178^{+0.077}_{-0.076}$	"Transit photometry"
Semi-major axis, a	au	0.0330±0.0002	"Transit photometry"
Impact parameter, b	–	$0.027^{+0.13}_{-0.023}$	"Transit photometry"
Quadratic limb-darkening coefficient u_1	–	0.393	"Transit photometry"
Quadratic limb-darkening coefficient u_2	–	0.219	"Transit photometry"
Combined parameters			
Semi-amplitude of the planet RVs, K_p	km s^{-1}	196.52±0.94	This work
Planet radius, R_p	$R_{♃}$	$1.854^{+0.077}_{-0.076}$	"Transit photometry"
Planet density, ρ_p	g cm^{-3}	0.17±0.02	This work
Planet surface gravity, g_p	m s^{-2}	6.4±0.5	This work
Total stellar irradiance	$\mathcal{S}_\odot^N$	4,104±896	This work
Equilibrium temperature for null albedo	K	2,228±122	This work
Dayside brightness temperature at 3.6 µm	K	2,693±56	Ref. [25]
Atmospheric scale height (dayside)	km	1,501±130	This work
Differential transit depth of one scale height	ppm	266±26	This work

In the rightmost column, text in quotes indicates named Methods sections in this paper (sometimes abbreviated); ref. [25] is also mentioned.

Extended Data Table 2 | Radial velocities of WASP-76 obtained with ESPRESSO

BJD	RV (m s^{-1})	σ_{RV} (m s^{-1})	Epoch
58364.65995	-1114.88	1.67	1
58364.66760	-1116.34	1.39	1
58364.67561	-1123.08	1.00	1
58364.68299	-1124.66	0.79	1
58364.69069	-1131.50	0.85	1
58364.86077	-1194.98	0.55	1
58364.86842	-1198.83	0.56	1
58364.87606	-1200.42	0.58	1
58364.88374	-1200.96	0.72	1
58364.89110	-1201.34	1.00	1
58364.89953	-1206.00	0.97	1
58364.90687	-1209.20	0.91	1
58364.91482	-1212.37	0.86	1
58364.92248	-1215.44	0.83	1
58422.55814	-1114.88	1.04	2
58422.56232	-1115.76	1.01	2
58422.56641	-1118.47	0.99	2
58422.57082	-1119.13	1.01	2
58422.57510	-1122.64	0.95	2
58422.57924	-1124.56	0.90	2
58422.58352	-1126.14	0.91	2
58422.58780	-1126.88	0.90	2
58422.59203	-1126.98	0.90	2
58422.59621	-1130.71	0.94	2
58422.60051	-1131.66	1.02	2
58422.60477	-1134.06	0.95	2
58422.60907	-1135.32	0.93	2
58422.61325	-1136.36	0.86	2
58422.77451	-1201.85	0.88	2
58422.77879	-1201.62	0.85	2
58422.78295	-1201.66	0.86	2
58422.78722	-1208.19	0.92	2
58422.79147	-1206.84	0.85	2
58422.79571	-1209.68	0.84	2
58422.79999	-1211.27	0.83	2
58422.80421	-1210.52	0.85	2
58422.80857	-1216.66	0.89	2
58422.81266	-1213.72	0.89	2
58422.81693	-1218.75	0.94	2
58422.82116	-1216.22	0.96	2
58422.82543	-1222.85	0.95	2
58422.82970	-1218.60	0.98	2
58422.83398	-1223.08	0.90	2
58422.83819	-1222.37	0.90	2
58422.84241	-1222.71	0.94	2
58422.84661	-1224.40	1.01	2
58684.91813	-1095.81	0.69	3
58684.92351	-1096.32	0.71	3
58695.85571	-1119.54	1.01	3
58696.92344	-1152.16	1.50	3
58706.93572	-1208.77	0.77	3
58719.85812	-1283.64	0.63	3
58721.85408	-1285.71	0.72	3
58725.88068	-1157.62	1.02	3
58731.80455	-1061.76	0.69	3
58741.64369	-1284.85	0.64	3
58741.65148	-1286.11	0.60	3
58741.65905	-1286.98	0.58	3
58741.66666	-1286.84	0.56	3
58752.63975	-1282.47	5.69	3
58752.64642	-1279.18	1.81	3
58752.65049	-1276.87	1.79	3
58752.65472	-1275.84	1.77	3
58752.65877	-1279.07	1.94	3
58752.66323	-1275.61	1.84	3
58752.66730	-1273.82	1.75	3
58752.67154	-1275.40	1.87	3
58752.67574	-1270.26	2.06	3
58752.67983	-1273.41	1.83	3
58752.68687	-1272.35	1.19	3
58752.69488	-1268.70	0.99	3
58752.70012	-1253.64	21.49	3
58753.67498	-1086.73	0.95	3
58754.79119	-1177.80	0.70	3

Offsets 1, 2 and 3 have been applied to epochs 1 (2018 September 02), 2 (2018 October 30) and 3 (autumn 2019), respectively. These data exclude the points obtained during transit.

Extended Data Table 3 | Orbital elements from the MCMC retrieval on the radial velocities

Parameter	Unit	Prior	Maximum likelihood	Posterior median
Period P	days	Gaussian (μ=1.80988198 d, σ=6.4×10^{-7} d)	1.8098821	1.8098819(7)
Semi-amplitude $K_\star$	m s^{-1}	Uniform on [0,30] km s^{-1}	115.94	$116.02^{+1.29}_{-1.35}$
$T_{\rm conj} - T_c$	s	Gaussian on $T_{\rm conj}$ (μ=58,080.626165 BJD, σ=4.1×10^{-4} d)	37	5^{+24}_{-34}
$\sqrt{e}\cos\omega$	–	Gaussian on e cos ω (μ=−0.0013, σ=8×10^{-4})	−0.0562	$-0.0169^{+0.0132}_{-0.0102}$
$\sqrt{e}\sin\omega$	–	Uniform	0.001	$-0.062^{+0.092}_{-0.078}$
σ_W^2	m s^{-1}	Truncated Gaussian on σ_W^2 (σ=100 m^2 s^{-2})	1.67	$1.29^{+0.25}_{-0.28}$
σ_R^2	m s^{-1}	Truncated Gaussian on σ_R^2 (σ=100 m^2 s^{-2})	1.41	$3.13^{+1.01}_{-1.33}$
Correlation time scale τ	days	Log-uniform on 1/τ on [0.001,1000] d	0.04	$1.090^{+2.56}_{-1.08}$
Offset 1	m s^{-1}	Uniform on [−200,200] km s^{-1}	−1,160.70	$-1{,}162.00^{+2.86}_{-2.63}$
Offset 2	m s^{-1}	Gaussian (μ=offset 1, σ=10 m s^{-1})	−1,167.78	$-1{,}167.54^{+2.79}_{-2.73}$
Offset 3	m s^{-1}	Gaussian (μ=offset 1, σ=10 m s^{-1})	−1,171.36	$-1{,}171.11^{+1.28}_{-1.36}$
Planet mass M_p	$M_{\oplus}$		0.894	$0.894^{+0.014}_{-0.013}$

Exploring dynamical phase transitions with cold atoms in an optical cavity

https://doi.org/10.1038/s41586-020-2224-x

Received: 27 September 2019

Accepted: 10 February 2020

Published online: 29 April 2020

Check for updates

Juan A. Muniz[1,3], Diego Barberena[1,2,3], Robert J. Lewis-Swan[1,2,3], Dylan J. Young[1], Julia R. K. Cline[1], Ana Maria Rey[1,2]✉ & James K. Thompson[1]✉

Interactions between atoms and light in optical cavities provide a means of investigating collective (many-body) quantum physics in controlled environments. Such ensembles of atoms in cavities have been proposed for studying collective quantum spin models, where the atomic internal levels mimic a spin degree of freedom and interact through long-range interactions tunable by changing the cavity parameters[1–4]. Non-classical steady-state phases arising from the interplay between atom–light interactions and dissipation of light from the cavity have previously been investigated[5–11]. These systems also offer the opportunity to study dynamical phases of matter that are precluded from existence at equilibrium but can be stabilized by driving a system out of equilibrium[12–16], as demonstrated by recent experiments[17–22]. These phases can also display universal behaviours akin to standard equilibrium phase transitions[8,23,24]. Here, we use an ensemble of about a million strontium-88 atoms in an optical cavity to simulate a collective Lipkin–Meshkov–Glick model[25,26], an iconic model in quantum magnetism, and report the observation of distinct dynamical phases of matter in this system. Our system allows us to probe the dependence of dynamical phase transitions on system size, initial state and other parameters. These observations can be linked to similar dynamical phases in related systems, including the Josephson effect in superfluid helium[27], or coupled atomic[28] and solid-state polariton[29] condensates. The system itself offers potential for generation of metrologically useful entangled states in optical transitions, which could permit quantum enhancement in state-of-the-art atomic clocks[30,31].

Arrays of ultracold alkaline-earth atoms with narrow-linewidth optical transitions are the basis of the most precise atomic clocks[31] and are also used for quantum simulation[32] and quantum information processing[33]. When these atoms are placed inside an optical cavity, their long-lived internal levels make them ideal to simulate non-equilibrium quantum magnetism, including models featuring long-range interactions mediated by cavity photons.

Here we report an advance towards the goal of simulating quantum magnetism in an optical cavity. We observe a dynamical phase transition generated by coupling a narrow-linewidth optical transition of an ensemble of strontium atoms to a single detuned cavity mode (Fig. 1a, left).

In general terms, non-equilibrium phase transitions, characterized by the existence of a critical point that separates phases with distinct properties, have been described in various contexts. In driven open systems, non-equilibrium phases are signalled by different steady states that depend on system parameters such as pump or loss rates[9–11,34–36], independent of initial conditions. Conversely, here we focus on a non-equilibrium phase transition in a closed system—often referred to as a dynamical phase transition (DPT)—where the non-equilibrium quantum phases are dynamical in nature: that is, qualitatively distinct behaviours are observed below, above or at a critical point[14,37–40] in terms of the time average of an order parameter such as magnetization. DPTs are typically initiated by quenching control parameters and depend on the initial state of the system. Such DPTs have been observed experimentally in arrays of trapped ions[17] and cold gases[20], as well as previously in the context of macroscopic self-trapping[28,29,41,42]. Here we demonstrate a DPT in a system of cold atoms with global interactions mediated by an optical cavity.

Implementation of the Lipkin–Meshkov–Glick model

A feature of our cavity simulator (Fig. 1a), compared with earlier observations, is the use of a much larger ensemble of $N \approx 10^5$–10^6 cold ^{88}Sr atoms. We use two long-lived electronic levels in these atoms, $|\downarrow\rangle$ ($^1S_0(m_J = 0)$) and $|\uparrow\rangle$ ($^3P_1(m_J = 0)$) states, to mimic a spin-1/2 system ($|\downarrow\rangle$ and $|\uparrow\rangle$, respectively). The atoms are confined in a one-dimensional (1D) optical lattice with a near-magic-wavelength of 813 nm supported by the optical cavity. We operate the experiment

[1]JILA, NIST and Department of Physics, University of Colorado, Boulder, CO, USA. [2]Center for Theory of Quantum Matter, University of Colorado, Boulder, CO, USA. [3]These authors contributed equally: Juan A. Muniz, Diego Barberena, Robert J. Lewis-Swan. ✉e-mail: arey@jila.colorado.edu; jkt@jila.colorado.edu

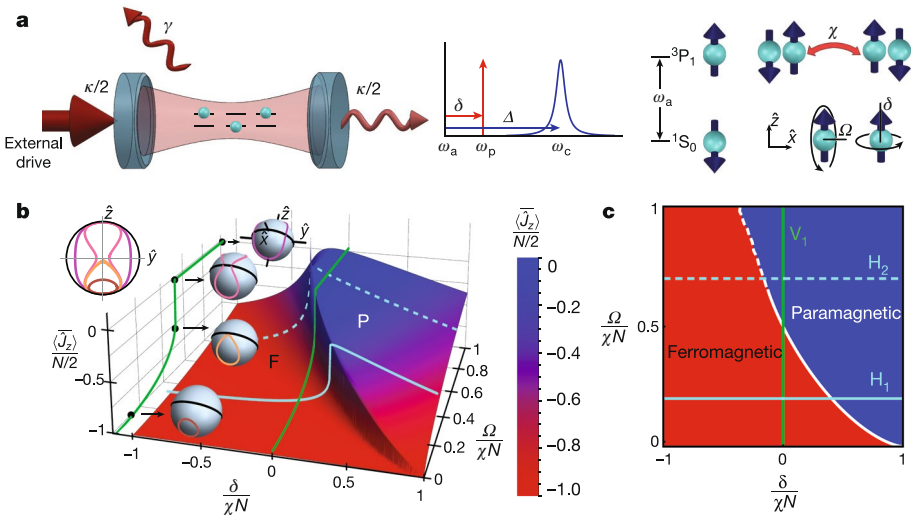


Fig. 1 | System and dynamical phase diagram. a, An ensemble of ^{88}Sr atoms is trapped in a 1D optical lattice supported by an optical cavity (left). The atoms are coupled to a single cavity mode with a single-photon Rabi frequency $2g$ and a resonance frequency ω_c detuned by $\Delta = \omega_c - \omega_a$ from the optical atomic transition $^1S_0$, $m_j = 0$ ($|\downarrow\rangle$) to $^3P_1$, $m_j = 0$ ($|\uparrow\rangle$) (with frequency ω_a and linewidth γ). Light leaks out of the cavity at a total rate κ. The cavity is driven externally by a laser with frequency ω_p that, if on resonance with an empty cavity, would establish a coherent state inside the cavity with average intracavity photon number $|2\Omega_p/\kappa|^2$. As shown centre and right, for the far-detuned cavity system in consideration, the external drive generates a transverse field that drives Rabi flopping at frequency $\Omega = -2g\Omega_p/\Delta$. The external drive detuning $\delta = \omega_p - \omega_a$ establishes a longitudinal field. The detuned cavity field generates an effective spin exchange interaction of strength $\chi = -g^2/\Delta$ (red arrow, right). **b**, The collective LMG model with transverse and longitudinal fields features a second-order DPT between paramagnetic (P, blue) and ferromagnetic phases (F, red). The DPT is characterized by the long-time average of the collective magnetization $\langle\bar{\hat{J}}_z\rangle$, and its dynamics can be characterized by trajectories of the classical Bloch vector in the pseudospin Bloch sphere (see projection and associated sphere insets). For $\delta = 0$, in the paramagnetic phase the trajectories circumnavigate the Bloch sphere, whereas in the ferromagnetic phase the trajectories are trapped below the equator. **c**, The two-dimensional map shows the DPT indicated by a sharp change in $\langle\bar{\hat{J}}_z\rangle$ (white solid line) for $\delta/(\chi N) \geq -1/8$. The white dashed line ($\delta/(\chi N) < -1/8$) signals a smooth crossover between the two phases (see Methods). Curves for $\delta = 0$ (green solid line, V_1), $\Omega/(\chi N) = 0.2$ (blue solid line, H_1) and $\Omega/(\chi N) = 0.7$ (blue dashed line, H_2) are shown on both diagrams and experimentally investigated in Figs. 2b, 3a and 3b, respectively. The dependence of the transition point on both $\delta/(\chi N)$ and $\Omega/(\chi N)$ is investigated in Extended Data Fig. 2b.

in a regime in which the atoms couple to a single common transverse electromagnetic (TEM_{00}) mode of the optical cavity with resonance frequency ω_c detuned by $\Delta = \omega_c - \omega_a$ from the atomic optical transition with frequency ω_a. Here $|\Delta|$ is large with respect to the linewidths of the cavity, $\kappa/(2\pi) = 153.0(4)$ kHz, and atomic transition, $\gamma/(2\pi) = 7.5$ kHz, and also the vacuum Rabi splitting $g\sqrt{N}$ induced by the atoms, with g the single-photon Rabi frequency $2g/(2\pi) = 21.8$ kHz. This means that the cavity-mediated dynamics of the atoms essentially conserves energy and can be well described by the following Hamiltonian:

$$\hat{H} = \hbar\chi\hat{J}^+\hat{J}^- + \hbar\Omega\hat{J}_x - \hbar\delta\hat{J}_z \quad (1)$$

This Hamiltonian can be recast as the well known Lipkin–Meshkov–Glick (LMG) model[25,26], which has been studied in various contexts, including quantum magnetism. In equation (1), we have introduced the collective spin operators $\hat{J}_\alpha = \sum_j \hat{\sigma}_j^\alpha/2$, where $\hat{\sigma}_j^\alpha$ is a Pauli operator for the jth atom with $\alpha = x, y, z$ and $\hat{J}^\pm = \hat{J}_x \pm i\hat{J}_y$. The summation runs over the individual atoms $j = 1,\ldots, N$ in the cavity. The parameter χ sets the strength of the infinite-range exchange interactions mediated by the cavity mode, and Ω and δ define the strength of the transverse and longitudinal fields respectively (Fig. 1a). The model is realized in the limit in which the cavity field couples identically to all atoms trapped in the optical lattice (see Methods for modifications due to inhomogeneity in this coupling).

Dynamical phase diagram of the LMG model

On varying the ratios between Ω, δ and χ, two distinct dynamical phases emerge, for which the time-averaged collective magnetization (along $\hat{z}$) of the atomic ensemble $\langle\bar{\hat{J}}_z\rangle \equiv \lim_{T\to\infty}(1/T)\int_0^T\langle\hat{J}_z(t)\rangle dt$ serves as an order parameter. When all spins are initially prepared in the $|\downarrow\rangle$ state and $\delta = 0$, the system features a sharp second-order transition[43] between a dynamical ferromagnetic phase with $\langle\bar{\hat{J}}_z\rangle \neq 0$ and a dynamical paramagnetic phase with $\langle\bar{\hat{J}}_z\rangle = 0$. This transition is indicated by the solid green line (V_1) on the phase diagram shown in Fig. 1b, as well as its projection on the $\langle\bar{\hat{J}}_z\rangle$ versus Ω plane in the same panel and in Fig. 1c. More generally, as a function of the parameters Ω and $\delta/(\chi N) \geq -1/8$, we observe a non-analyticity of the order parameter $\langle\bar{\hat{J}}_z\rangle$ (indicated by a solid white line in Fig. 1c), which marks a second-order transition between the two dynamical phases. However, the transition line is interrupted at a critical point $\delta/(\chi N) = -1/8$. Beyond this, there is a smooth crossover regime (indicated by a white dashed line in Fig. 1c) in which the system is ruled mainly by single-particle physics (set by δ and Ω) and has an intermediate behaviour between that of a ferromagnet and a paramagnet.

In the ferromagnetic phase (red region in Fig. 1b and c), the instantaneous magnetization $\langle\hat{J}_z\rangle$ oscillates about a non-zero time-averaged value, and the collective pseudospin Bloch vector $\langle\hat{\mathbf{J}}\rangle \equiv (\langle\hat{J}_x\rangle, \langle\hat{J}_y\rangle, \langle\hat{J}_z\rangle)$ remains trapped below the equator of the Bloch sphere throughout the dynamics. This phase is dominated by the interactions which can be understood in a mean-field approximation as $\chi\hat{J}^+\hat{J}^- \approx \chi(\hat{\mathbf{J}}\cdot\hat{\mathbf{J}} - \hat{J}_z^2) \approx \chi(N/2)(N/2+1) - 2\chi\langle\hat{J}_z\rangle\hat{J}_z$. The term $\hat{\mathbf{J}}\cdot\hat{\mathbf{J}}$ is a constant when restricted to the fully symmetric spin manifold, which is the case of interest here. The second term describes a self-induced precession of the collective Bloch vector about the $\hat{z}$ axis, which effectively tilts the axis of rotation of the comparatively weak transverse field, such that the trajectory of the Bloch vector deforms into an orbit that remains below the equatorial plane.

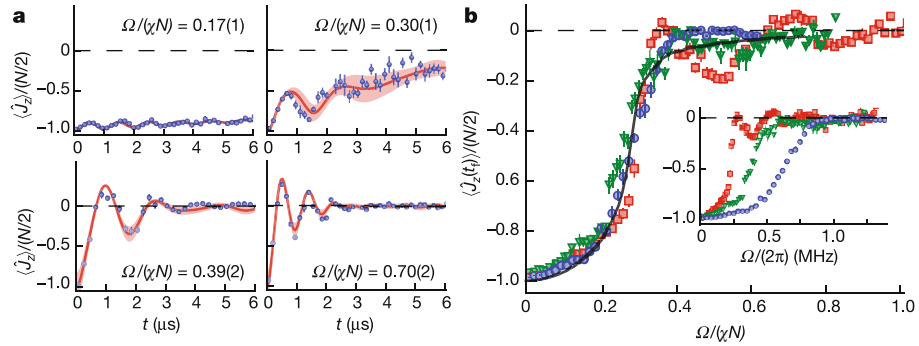


Fig. 2 | Characteristic evolution of dynamical phases and scaling of DPT with atom number. a, Time-traces of the mean magnetization $\langle \hat{J}_z \rangle$ for the case of all spins initially on $|\downarrow\rangle$, $\delta = 0$ and Ω quenched to different values, at $t = 0$, in the ferromagnetic (top panels) and paramagnetic (bottom panels) phases for $N = 950 \times 10^3$ atoms and $\Delta/(2\pi) = 50$ MHz. The experimental data (blue) are compared to theoretical calculations (red lines) based on a mean-field description including relevant experimental details (see Methods and Supplementary Information). Shaded theoretical region accounts for shot-to-shot fluctuations in $\Omega/(\chi N)$. Each point is the average of 12 experimental repetitions. **b**, Magnetization $\langle \hat{J}_z(t_f) \rangle$ for different numbers of atoms $N = (935, 620, 320) \times 10^3$ (blue, green and red, respectively) after $t_f = 4$ μs of evolution for different normalized drive strengths $\Omega/(\chi N)$ for $\Delta/(2\pi) = 50$ MHz and $\delta = 0$. This measurement maps to the green solid line (V_1) in Fig. 1b and c. The drive-strength normalization in each experimental shot is done by spin-dependent imaging. The solid black line indicates the simulated average (0–6 μs) as a function of the normalized drive including dephasing sources. The inset shows the magnetization versus non-normalized transverse field strength Ω for the same data sets. All error bars in experimental data are statistical (1σ).

Conversely, the paramagnetic phase (blue region in Fig. 1b and c) is dominated by Rabi flopping driven by the transverse field $\Omega \hat{J}_x$. This term causes large oscillations of the instantaneous $\langle \hat{J}_z \rangle$, and, for $\delta = 0$, the collective Bloch vector breaks through the equatorial plane and rotates about the entire Bloch sphere.

For $\delta = 0$, the transition between the paramagnetic and ferromagnetic phases occurs at a critical drive $\Omega_c = \chi N/2$, as shown in Fig. 1b and c. The sharp transition in the dynamical behaviour of the system is traced back to the change in direction of the self-generated precession proportional to $2\chi \langle \hat{J}_z \rangle \hat{J}_z$ as the Bloch vector crosses the equatorial plane at $\langle \hat{J}_z \rangle = 0$, generating an abrupt shift to large-amplitude oscillations for $\Omega > \Omega_c$. Typical dynamics of the collective Bloch vector in the ferromagnetic and paramagnetic phases are shown as insets in Fig. 1b. The solid green, solid blue and dashed blue lines in Fig. 1b and c indicate analogous trajectories in the phase diagram which will be explored experimentally later in Figs. 2b, 3a and 3b, respectively (see also Extended Data Fig. 2b and Supplementary Fig. 3 for investigation of the transition as a function of detuning and drive).

Probing the LMG dynamical phase diagram

In our simulator, the cavity mediates a global spin-exchange interaction, which is microscopically described by a flip-flop process in which the emission of a photon from atom i in state $|\uparrow\rangle$ into the cavity mode is subsequently absorbed by atom j in state $|\downarrow\rangle$ (Fig. 1a, right). We operate in the regime $|\Delta| \gg g\sqrt{N}$, where the instantaneous average number of photons in the cavity mediating the interaction is much less than N, and the dynamics are well described by a spin-exchange model $\chi \hat{J}^+ \hat{J}^-$ with coupling constant $\chi = -g^2/\Delta$ (see also Extended Data Fig. 2a and Methods). Similarly, the large detuning means that superradiant emission does not play an active role, in contrast to previous work[2]. The interaction dynamics are faster than spontaneous emission, $|\chi|N \gg \gamma$, and satisfy the hierarchy $|\Delta| \gg g\sqrt{N} \gg \kappa, \gamma$.

We realize the transverse fields Ω and δ by injecting laser light at frequency ω_p into the optical cavity through one mirror, creating a coherent driving field $\Omega_p e^{i\omega_p t}$ inside the cavity. In the rotating frame at ω_p, the laser light's detuning from atomic resonance $\delta = \omega_p - \omega_a$ provides the longitudinal field $\delta \hat{J}_z$ in equation (1). Moreover, the applied laser rapidly builds up a classical field within the cavity on a timescale of approximately $1/\Delta$, which couples $|\downarrow\rangle$ to $|\uparrow\rangle$. This realizes the transverse field $\Omega \hat{J}_x$ in equation (1), where $\Omega = -2g\Omega_p/\Delta$. We adopt the convention that this transverse field is oriented along $\hat{x}$ in the pseudospin coordinate system such that by jumping the phase of the laser light, we are able to create transverse fields oriented along any direction in the pseudospin x–y plane. Furthermore, the experiment is realized with a standing wave cavity, where incommensurate lattice and drive wavelengths generate inhomogeneous Ω and χ parameters compared with the ideal case presented above. This leaves unchanged the generic features of the phase diagram in Fig. 1, but quantitatively modifies the phase boundary.

DPT in absence of longitudinal field

In Fig. 2, we show experimental observations of the characteristic dynamics and DPT. We begin with all atoms in $|\downarrow\rangle$ and then quench Ω from zero to a specific value at $t = 0$. After a variable evolution time, we rapidly freeze the atomic dynamics by quenching $\Omega \to 0$ and creating strong single-particle dephasing of the ground state. The atomic magnetization $\langle \hat{J}_z \rangle$ and atom number N are then measured with high efficiency using fluorescence in combination with electron shelving and state-dependent displacements (see Methods and Extended Data Fig. 1).

For the time traces presented in Fig. 2a, we map the magnetization across different drive strengths with fixed $\delta = 0$. For drives deep in the ferromagnetic phase (Fig. 2a, top left), we observe small-amplitude oscillations that are in excellent agreement with our theoretical model based on a mean-field description of the system (see Methods and Supplementary Information). Close to the experimental critical point (Fig. 2a, top right and bottom left), the dynamics become more complicated owing to the complex interplay between interactions, drive and single-particle decoherence due to undesirable atomic motion in the optical lattice (see Supplementary Information and Supplementary Figs. 1 and 2). Deep in the paramagnetic phase (Fig. 2a, bottom right), we observe dynamics of the magnetization consistent with single-particle Rabi flopping with frequency Ω and in good agreement with our simulation. Damping of the oscillations occurs predominantly because of inhomogeneity in the coupling of the spins to the common cavity mode, shot-to-shot fluctuations in $\Omega/(\chi N)$ (attributed mostly to atom number fluctuations at about the 5% (root mean square, r.m.s.) level) and atomic motion in the lattice. Spontaneous emission and decoherence related to leakage of photons from the cavity are negligible. We include these

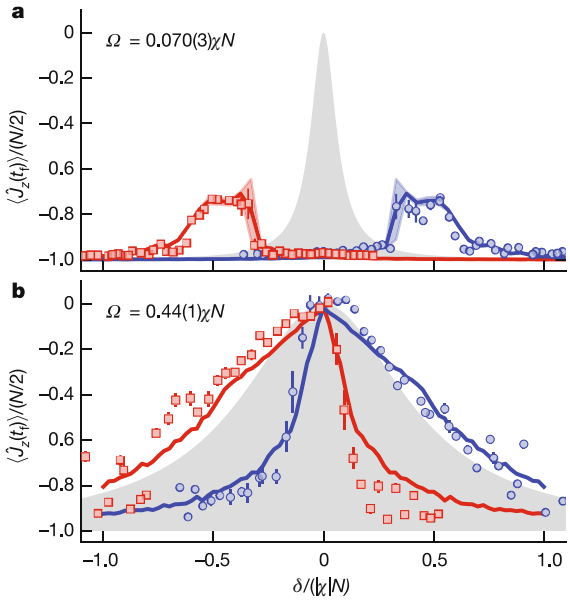


Fig. 3 | Characterization of the DPT as a function of longitudinal field for two different transverse field values at fixed χN. a, **b**, The atomic magnetization $\langle \hat{J}_z(t_f)\rangle$ at $t_f = 4$ μs is measured as a function of the normalized drive detuning $\delta/(|\chi|N)$ for cavity detunings $\Delta/(2\pi) = \pm 50$ MHz (red, +50 MHz; blue, −50 MHz) for two different drive strengths: (**a**) $\Omega = 0.070(3)\chi N$ and (**b**) $\Omega = 0.44(1)\chi N$. The inner edges of the resonant features in **a** indicate a sharp transition from ferromagnetic to paramagnetic phases as $|\delta|$ is increased. In contrast, the corresponding crossover in **b** is smoothed. Numerical simulations are shown as blue and red solid lines with corresponding shaded regions. The grey-shaded area indicates the non-interacting limit of Rabi flopping. Measurements in **a** and **b** map, respectively, to cuts represented by the blue lines H_1 and H_2 in Fig. 1b and c. All error bars in experimental data are statistical (1σ).

effects in our theoretical model (Fig. 2a, red solid line), and fluctuations in $\Omega/(\chi N)$ are indicated by the red shaded regions. Typically, we notice that the experimentally calibrated parameters overestimate the value of $\Omega/(\chi N)$ by about 10% compared with the numerical simulations. We attribute this systematic disagreement to drifts on the calibration parameters and to details not captured by the theoretical model (see Supplementary Information).

We characterize the behaviour of the DPT with system size by measuring $\langle \hat{J}_z(t_f)\rangle$ at time $t_f = 4$ μs for different atom number N while initializing every atom in $|\downarrow\rangle$ for $\delta = 0$ in Fig. 2b. Measuring $\langle \hat{J}_z\rangle$ at a fixed time serves as a proxy of the long-term time-averaged magnetization, as considerable damping is caused by the previously mentioned effects. In the Fig. 2b inset, we observe a transition in the magnetization at different values of the transverse field Ω, depending on atom number N. The dependence of the transition as a function of system size is demonstrated by re-scaling the corresponding drive as $\Omega/(\chi N)$, as shown in the main panel of Fig. 2b, analogous to the green curve in Fig. 1b and c. We observe collapse of the data and a critical drive $\Omega_c^{exp} = 0.35(3)\chi N$. A comparison to theoretical calculations using time-averaged magnetization (see Methods) shows reasonable agreement (solid black line). The shift of the critical point relative to the ideal collective model, $\Omega_c/(\chi N) = 1/2$, is predominantly attributable to the spatial inhomogeneity in the coupling of the atoms to the cavity mode (see Methods). Other small factors include single-particle decoherence of the atoms, which also contributes to the smearing out of the sharpness of the transition observed in the ideal system (Fig. 1b). Nevertheless, a clear transition can be observed, as shown by comparing to the theoretical calculation.

DPTs at fixed transverse fields

The DPT can also be probed using our ability to controllably introduce a longitudinal field proportional to $\delta \hat{J}_z$ by detuning the injected light from the atomic transition, as shown in Fig. 1b. In Fig. 3, we map out the response of the system to the drive detuning δ by measuring the order parameter $\langle \hat{J}_z(t_f)\rangle$ at $t_f = 4$ μs for two fixed values of the drive strength Ω above and below the ($\delta = 0$) critical point, Ω_c^{exp}, and for two opposite cavity detunings $\Delta/(2\pi) = \pm 50$ MHz.

We observe a sharp transition in the order parameter $\langle \hat{J}_z\rangle$ versus drive detuning, separating the ferromagnetic and the paramagnetic dynamical phases for a drive below the observed critical point Ω_c^{exp} (blue solid line, H_1, in Fig. 1b and c). This is plotted in Fig. 3a with $\Omega = 0.070(3)\chi N < \Omega_c^{exp}$. We observe sharp transitions at the inside edges of the resonant features, which occur symmetrically for each Δ at $\delta_c/|\chi N| = \mp 0.27(2)$. The critical value of δ_c and the gradual decrease in $\langle \hat{J}_z\rangle$ for large detuning show good agreement with a mean-field calculation. The robustness of the sharp transition is demonstrated by the symmetric response of the magnetization for $\Delta \leftrightarrow -\Delta$ and thus of the interaction shift $\chi N \leftrightarrow -\chi N$.

Conversely, when the drive is tuned above Ω_c^{exp}, $\Omega = 0.44(1)\chi N > \Omega_c^{exp}$ (Fig. 3b), indicated by the blue dashed line, H_2, in Fig. 1b and c, we observe a smoother crossover between the paramagnetic and ferromagnetic phases about the detuning $\delta_c/|\chi N| = \pm 0.04(3)$ in agreement with the mean-field calculation. Tuning $\delta < \delta_c$ ($\Delta > 0$) reduces the influence of the collective interactions, and the magnetization resembles the prediction of single-particle detuned Rabi flopping.

In both cases, the response of the system to δ can be understood by interpreting the single-particle shift and interaction in equation (1) as a nonlinear detuning proportional to $(2\chi\langle \hat{J}_z\rangle + \delta)\hat{J}_z$, which competes with the coherent drive. Depending on the sign of the interaction and the instantaneous magnetization, the single-particle term δ can either cancel or enhance the contribution of the interactions relative to the coherent drive, tuning the system between the ferromagnetic and paramagnetic dynamical phases. The predominant role of the interactions in the dynamics, especially below the critical point, can be observed by contrasting with the purely single-particle model of detuned Rabi oscillations (grey shaded area), which predicts a Lorentzian lineshape centred at $\delta = 0$.

Sensitivity to initial condition

The single-particle control achievable in our experimental platform allows us to explore the DPT as a function of the initial state, as shown in Fig. 4. Specifically, we are able to demonstrate that the critical point of the transition is state-dependent, by preparing the collective pseudospin in different positions on the Bloch sphere. For example, we can prepare the system with $\Omega < \Omega_c^{exp}$ such that the initial collective states near the south pole remained trapped below the equator, yet there also exist initial states prepared further towards the equator that exhibit large oscillations around the Bloch sphere characteristic of the paramagnetic phase.

Probing the response of the dynamics to different initial conditions allows us to establish a connection between the DPT in our effective spin model and the phenomena of macroscopic self-trapping and Josephson tunnelling observed in coupled atomic condensates[28] and solid-state polariton condensates[29]. Figure 4a schematically shows a double-well atomic condensate, where the initial magnetization of the collective state on the Bloch sphere is analogous to the initial population imbalance between the wells, while the azimuthal angle maps to the relative phase difference of the condensates. Similarly, the ferromagnetic and paramagnetic phases can be related to the self-trapped and tunnelling phases respectively[41].

In Fig. 4c, we plot the measured magnetization after 4 μs of evolution a polar projection of the Bloch sphere for different drive strengths Ω,

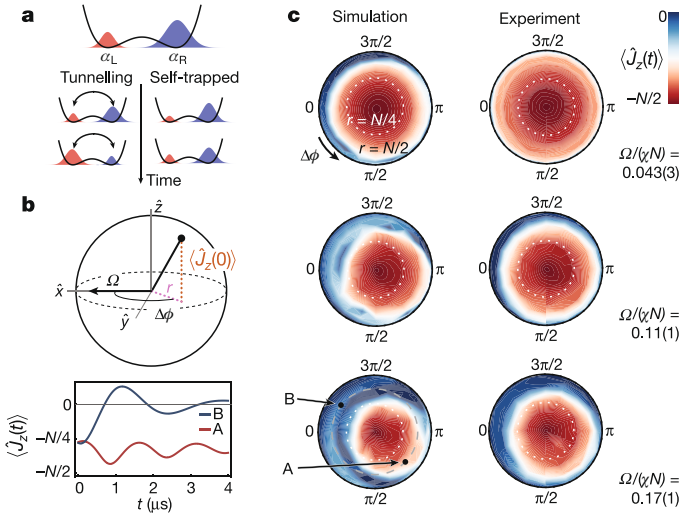


Fig. 4 | Dependence of dynamical phases on initial conditions. a, The initial state on the Bloch sphere and subsequent dynamics of the spin model can be mapped to that of atomic condensates in a double-well potential, described by coherent complex amplitudes in the left and right wells, α_L and α_R, respectively. Population imbalance maps to the magnetization ($J_z \approx |\alpha_L|^2 - |\alpha_R|^2$) and the relative phase of the condensate wavefunctions maps to the azimuthal angle of the spin state ($\Delta\phi$). As time evolves, the population imbalance either oscillates as atoms tunnel back and forth between the wells (tunnelling phase) or remains approximately constant (self-trapped phase). **b**, The initially prepared spin state can be parameterized in terms of the projection onto the equatorial plane $r = \sqrt{(N/2)^2 - \langle J_z(0)\rangle^2}$ and the relative azimuthal phase $\Delta\phi$ between the initial collective Bloch vector and the coherent drive (top; see Methods). The plot of $\langle \hat{J}_z(t)\rangle$ (bottom) indicates typical dynamics in the red and blue regions shown in the adjacent panel of **c**. **c**, Colour map of $\langle J_z(t_f)\rangle$ at $t_f = 4$ μs of evolution, plotted in a polar projection of the Bloch sphere with coordinates defined by the initial condition as in **b** (initial conditions are always below the equator; they are shown above the equator in the figure to simplify visualization). Left (right) panels show simulated (experimental) results for $\langle J_z(t_f)\rangle$ at $t_f = 4$ μs of evolution for different normalized drives $\Omega/(\chi N)$.

as we scan the initially prepared state $\mathbf{J}(0)$. Here, the radial coordinate maps to the magnitude of the projection of $\mathbf{J}(0)$ on the equatorial plane (for $\langle \hat{J}_z(0)\rangle < 0$), and the angle $\Delta\phi$ maps to the relative phase between the coherent drive and $\mathbf{J}(0)$ (see Methods). As we increase the drive strength, the set of initial conditions that lead to the ferromagnetic phase shrinks (red region) while also becoming increasingly asymmetric about the south pole. Both of these features are in qualitative agreement with our theoretical calculations, also shown in Fig. 4c, which take into account coupling inhomogeneities, dephasing and shot-to-shot fluctuations on $\Omega/(\chi N)$. Quantitative differences are predominantly due to neglecting axial motion of the atoms in the theoretical model.

Conclusion

The demonstration of the cooperation and competition between coherent drive and infinite-range interactions in an optical transition opens a path to the quantum simulation of richer spin models and out-of-equilibrium physics. For example, more complex spin–spin couplings can be engineered by using the available Zeeman sublevels of the $^3P_1$ state with two different cavity polarizations[3]. Moreover, in the presence of additional inhomogeneous terms, our system can explore dynamical phases predicted to exist in Bardeen–Cooper–Schrieffer superconductors[44,45], and by modulation of the transverse field our platform should be able to realize the archetypal model of a kicked top[46], relevant for explorations of quantum chaos and scrambling dynamics[47]. Lastly, our investigation of non-equilibrium dynamics with the ^{88}Sr ($^1S_0$–$^3P_1$) optical transition can lead to insight into how to generate entangled states for quantum sensing with the long-lived ^{87}Sr ($^1S_0$–$^3P_0$) optical transition used in state-of-the-art atomic clocks[30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2224-x.

1. Leroux, I. D., Schleier-Smith, M. H. & Vuletić, V. Implementation of cavity squeezing of a collective atomic spin. *Phys. Rev. Lett.* **104**, 073602 (2010).
2. Norcia, M. A. et al. Cavity-mediated collective spin-exchange interactions in a strontium superradiant laser. *Science* **361**, 259–262 (2018).
3. Davis, E. J., Bentsen, G., Homeier, L., Li, T. & Schleier-Smith, M. H. Photon-mediated spin-exchange dynamics of spin-1 atoms. *Phys. Rev. Lett.* **122**, 010405 (2019).
4. Vaidya, V. D. et al. Tunable-range, photon-mediated atomic interactions in multimode cavity QED. *Phys. Rev. X* **8**, 011002 (2018).
5. Baumann, K., Guerlin, C., Brennecke, F. & Esslinger, T. Dicke quantum phase transition with a superfluid gas in an optical cavity. *Nature* **464**, 1301–1306 (2010).
6. Klinder, J., Keßler, H., Wolke, M., Mathey, L. & Hemmerich, A. Dynamical phase transition in the open Dicke model. *Proc. Natl Acad. Sci. USA* **112**, 3290–3295 (2015).
7. Baden, M. P., Arnold, K. J., Grimsmo, A. L., Parkins, S. & Barrett, M. D. Realization of the Dicke model using cavity-assisted Raman transitions. *Phys. Rev. Lett.* **113**, 020408 (2014).
8. Ritsch, H., Domokos, P., Brennecke, F. & Esslinger, T. Cold atoms in cavity-generated dynamical optical potentials. *Rev. Mod. Phys.* **85**, 553–601 (2013).
9. Landini, M. et al. Formation of a spin texture in a quantum gas coupled to a cavity. *Phys. Rev. Lett.* **120**, 223602 (2018).
10. Kroeze, R. M., Guo, Y., Vaidya, V. D., Keeling, J. & Lev, B. L. Spinor self-ordering of a quantum gas in a cavity. *Phys. Rev. Lett.* **121**, 163601 (2018).
11. Kroeze, R. M., Guo, Y. & Lev, B. L. Dynamical spin–orbit coupling of a quantum gas. *Phys. Rev. Lett.* **123**, 160404 (2019).
12. Heyl, M., Polkovnikov, A. & Kehrein, S. Dynamical quantum phase transitions in the transverse-field Ising model. *Phys. Rev. Lett.* **110**, 135704 (2013).
13. Žunkovič, B., Heyl, M., Knap, M. & Silva, A. Dynamical quantum phase transitions in spin chains with long-range interactions: merging different concepts of nonequilibrium criticality. *Phys. Rev. Lett.* **120**, 130601 (2018).
14. Eckstein, M., Kollar, M. & Werner, P. Thermalization after an interaction quench in the Hubbard model. *Phys. Rev. Lett.* **103**, 056403 (2009).
15. Lamacraft, A. & Moore, J. in *Ultracold Bosonic and Fermionic Gases* (eds Levin, K. et al.) 177–202 (Elsevier, 2012).
16. Nandkishore, R. & Huse, D. A. Many-body localization and thermalization in quantum statistical mechanics. *Annu. Rev. Condens. Matter Phys.* **6**, 15–38 (2015).
17. Zhang, J. et al. Observation of a many-body dynamical phase transition with a 53-qubit quantum simulator. *Nature* **551**, 601–604 (2017).
18. Fläschner, N. et al. Observation of dynamical vortices after quenches in a system with topology. *Nat. Phys.* **14**, 265 (2018).
19. Jurcevic, P. et al. Direct observation of dynamical quantum phase transitions in an interacting many-body system. *Phys. Rev. Lett.* **119**, 080501 (2017).
20. Smale, S. et al. Observation of a transition between dynamical phases in a quantum degenerate Fermi gas. *Sci. Adv.* **5**, eaax1568 (2019).
21. Zhang, J. et al. Observation of a discrete time crystal. *Nature* **543**, 217–220 (2017).
22. Choi, S. et al. Observation of discrete time-crystalline order in a disordered dipolar many-body system. *Nature* **543**, 221–225 (2017).
23. Prüfer, M. et al. Observation of universal dynamics in a spinor Bose gas far from equilibrium. *Nature* **563**, 217–220 (2018).
24. Erne, S., Bücker, R., Gasenzer, T., Berges, J. & Schmiedmayer, J. Universal dynamics in an isolated one-dimensional Bose gas far from equilibrium. *Nature* **563**, 225–229 (2018).
25. Lipkin, H., Meshkov, N. & Glick, A. Validity of many-body approximation methods for a solvable model: (I). Exact solutions and perturbation theory. *Nucl. Phys.* **62**, 188–198 (1965).
26. Ribeiro, P., Vidal, J. & Mosseri, R. Thermodynamical limit of the Lipkin–Meshkov–Glick model. *Phys. Rev. Lett.* **99**, 050402 (2007).
27. Backhaus, S. et al. Discovery of a metastable-state in a superfluid ^{3}He weak link. *Nature* **392**, 687–690 (1998).
28. Albiez, M. et al. Direct observation of tunneling and nonlinear self-trapping in a single bosonic Josephson junction. *Phys. Rev. Lett.* **95**, 010402 (2005).
29. Abbarchi, M. et al. Macroscopic quantum self-trapping and Josephson oscillations of exciton polaritons. *Nat. Phys.* **9**, 275–279 (2013).
30. Campbell, S. L. et al. A Fermi-degenerate three-dimensional optical lattice clock. *Science* **358**, 90–94 (2017).
31. Ludlow, A. D., Boyd, M. M., Ye, J., Peik, E. & Schmidt, P. O. Optical atomic clocks. *Rev. Mod. Phys.* **87**, 637–701 (2015).
32. Cazalilla, M. A. & Rey, A. M. Ultracold Fermi gases with emergent SU(n) symmetry. *Rep. Prog. Phys.* **77**, 124401 (2014).
33. Daley, A. J. Quantum computing and quantum simulation with group-II atoms. Quantum Inform. *Process.* **10**, 865 (2011).

34. Marino, J. & Diehl, S. Quantum dynamical field theory for nonequilibrium phase transitions in driven open systems. *Phys. Rev. B* **94**, 085150 (2016).
35. Barberena, D., Lewis-Swan, R. J., Thompson, J. K. & Rey, A. M. Driven-dissipative quantum dynamics in ultra-long-lived dipoles in an optical cavity. *Phys. Rev. A* **99**, 053411 (2019).
36. Mivehvar, F., Piazza, F. & Ritsch, H. Disorder-driven density and spin self-ordering of a Bose–Einstein condensate in a cavity. *Phys. Rev. Lett.* **119**, 063602 (2017).
37. Schiró, M. & Fabrizio, M. Time-dependent mean field theory for quench dynamics in correlated electron systems. *Phys. Rev. Lett.* **105**, 076401 (2010).
38. Sciolla, B. & Biroli, G. Quantum quenches and off-equilibrium dynamical transition in the infinite-dimensional Bose–Hubbard model. *Phys. Rev. Lett.* **105**, 220401 (2010).
39. Gambassi, A. & Calabrese, P. Quantum quenches as classical critical films. *Europhys. Lett.* **95**, 66007 (2011).
40. Smacchia, P., Knap, M., Demler, E. & Silva, A. Exploring dynamical phase transitions and prethermalization with quantum noise of excitations. *Phys. Rev. B* **91**, 205136 (2015).
41. Smerzi, A., Fantoni, S., Giovanazzi, S. & Shenoy, S. R. Quantum coherent atomic tunneling between two trapped Bose–Einstein condensates. *Phys. Rev. Lett.* **79**, 4950–4953 (1997).
42. Reinhard, A. et al. Self-trapping in an array of coupled 1D Bose gases. *Phys. Rev. Lett.* **110**, 033001 (2013).
43. Lerose, A., Žunkovič, B., Marino, J., Gambassi, A. & Silva, A. Impact of non-equilibrium fluctuations on pre-thermal dynamical phase transitions in long-range interacting spin chains. *Phys. Rev. B* **99**, 045128 (2019).
44. Barankov, R. A., Levitov, L. S. & Spivak, B. Z. Collective Rabi oscillations and solitons in a time-dependent BCS pairing problem. *Phys. Rev. Lett.* **93**, 160401 (2004).
45. Yuzbashyan, E. A., Dzero, M., Gurarie, V. & Foster, M. S. Quantum quench phase diagrams of an *s*-wave BCS–BEC condensate. *Phys. Rev. A* **91**, 033628 (2015).
46. Swingle, B., Bentsen, G., Schleier-Smith, M. & Hayden, P. Measuring the scrambling of quantum information. *Phys. Rev. A* **94**, 040302 (2016).
47. Swingle, B. Unscrambling the physics of out-of-time-order correlators. *Nat. Phys.* **14**, 988–990 (2018).

Methods

Experimental description

Our experiment begins by loading up to 10^6 ^{88}Sr atoms from a magneto-optical trap into a one-dimensional optical lattice within an optical cavity, as we have described previously[2,48–50]. The lattice has wavelength 813 nm. This lattice is nominally near-magic with respect to the ultra-narrow millihertz $^1S_0 \rightarrow {}^3P_0$ clock transition at 698 nm, but can be made near-magic-wavelength for our optical transition at 689 nm ($m_J = 0$ states), $^1S_0 \rightarrow {}^3P_1$, by setting the angle between the linear polarization of the lattice and the quantization axis. The near-magic wavelength reduces potential dephasing due to the transverse spreading of the atoms in a non-magic trap. We estimate residual inhomogeneous broadening due to the lattice to be below 2 kHz. The lattice spacing is incommensurate with the intracavity probe standing wave, leading to inhomogeneous coupling to the cavity mode. A sketch of the system is shown in Extended Data Fig. 1a. The atoms are laser-cooled to 14 μK and trapped in the optical lattice, with typical axial trap oscillation frequency $\omega_{trap}/(2\pi)$=200 kHz. The atom number is measured using fluorescence imaging on the dipole-allowed $^1S_0 \rightarrow {}^1P_1$ transition at 461 nm, and it is calibrated by comparing it with the vacuum Rabi splitting when the cavity is on resonance with the atomic transition ($\Delta = 0$), as detailed in ref. [48]. We determine $\Delta = 0$ and $\delta = 0$ from measurements of the symmetry of the collective vacuum Rabi splitting. The measured cavity linewidth is $\kappa/(2\pi) = 153.0(4)$ kHz. The cavity length is adjusted using piezoelectric actuators, such that it can be kept at a detuning Δ during the experiment.

The cavity is driven for a time τ by a near-resonant laser that realizes a coherent driving field $\Omega_p e^{i\omega_p t}$ in the cavity, as shown in Extended Data Fig. 1, where Ω_p is related to the input power P by the expression $\Omega_p = \sqrt{\kappa_m P/(2\hbar\omega_p)}$, with $\kappa_m = \kappa T_m/(T_m + T_L)$. Here, we define T_m and T_L as the single-mirror power transmission and loss coefficients, 105 ppm and 23 ppm respectively. The drive is turned on and off in approximately 10 ns using an in-fibre electro-optical modulator (EOM), which creates a sideband at detuning δ while other frequency components are far from resonance and suppressed by being even further from resonance with the cavity mode. We apply a strong magnetic field perpendicular to the cavity axis to define the quantization axis. The probe light is polarized along the magnetic field direction such that the system is an effective two-level system $|\downarrow\rangle = |{}^1S_0, m_J = 0\rangle$ and $|\uparrow\rangle = |{}^3P_1, m_J = 0\rangle$ transition. For a more complete energy level diagram, see Extended Data Fig. 1c).

To observe the DPT, we need to be able to take a snapshot of the magnetization $\langle \hat{J}_z \rangle$ after some period of dynamical evolution. To achieve this, we have developed a technique to freeze the dynamics quickly and then apply state-dependent spatial displacements of the cloud such that the populations in the ground and excited states $N_\downarrow$ and $N_\uparrow$ are imaged on two different regions of a charge-coupled device (CCD) (Extended Data Fig. 1b).

After the drive is applied for some time τ, as shown in the time sequence in Extended Data Fig. 1d, we turn off the coherent drive by extinguishing the applied EOM sideband. To effectively count atoms in both excited and ground state immediately after the drive, and freeze any dynamics that could be caused by spontaneous emission or the transient decay of the cavity field, we shine a strongly focused 461-nm beam along the $\hat{z}$ axis and apply a strong 688-nm shelving beam. The 461-nm beam immediately stops the dynamics as it dephases the atoms, overwhelming the single particle rotation and any collective interactions. In addition, the 461-nm beam exerts a radiation pressure force that gives a momentum kick to the ground-state atoms, causing them to move away from the trapping region. Simultaneously, the shelving beam optically pumps excited-state atoms to the metastable $^3P_{0,2}$ states (Extended Data Fig. 1c). We apply the shelving pulse for 5 μs. For scale, at 2 μs, we observe that >90% of the atoms in the excited state are shelved.

To finish our state-dependent detection, we allow for a short time of flight (about 100 μs) so that the momentum kick applied to the ground state atoms is translated into a displacement in space of a few 100 μm. We then optically pump the shelved atoms back to $^3P_1$ using 679-nm and 707-nm light applied for 200 μs. The atoms then decay to the ground state via single-atom decay with time constant 21 μs. We then perform fluorescence imaging for 50 μs to observe the number of atoms in the two spatially resolved clouds as shown in Extended Data Fig. 1b. This allows us to measure the magnetization $\langle \hat{J}_z \rangle = (N_\uparrow - N_\downarrow)/(2(N_\downarrow + N_\uparrow))$ and total atom number $N = N_\downarrow + N_\uparrow$ in a single shot. We found that the whole process efficiency is above 98%, limited mostly by the efficiency of the shelving process.

In Fig. 3, we change the drive detuning δ by changing the frequency of the radiofrequency pulse applied to the EOM. The grey shaded area represents the r.m.s. amplitude for Rabi oscillations without interactions, that is, $\chi = 0$ in our model, and the corresponding r.m.s. magnetization is calculated simply as

$$\langle \hat{J}_z \rangle^{\mathrm{rms}}_{\chi=0} = -\frac{N}{2}\frac{\Omega^2}{(\delta^2 + \Omega^2)} - \frac{N}{2} \tag{2}$$

In Fig. 4, the initial state preparation is accomplished by preparing each spin in $|\downarrow\rangle$ and then rotating the spins with a strong drive $\Omega > \Omega_c$ for some chosen time. At this point, $t = 0$, the system has acquired a magnetization $\langle \hat{J}_z(0) \rangle$. We then simultaneously shift the phase of the driving field by $\Delta\phi - \pi/2$ and its amplitude to some $\Omega < \Omega_c$ and evolve for a fixed time, typically 4 μs. The phase and amplitude jumps are accomplished by changing the phase and amplitude of the radiofrequency tone driving the EOM. We are then able to initialize the collective pseudospin Bloch vector at different positions on the Bloch sphere, such that $\langle \hat{J}_z(0) \rangle$ and $\Delta\phi$ define the polar and the azimuthal angles, respectively, as indicated in the figure in the main text. As the phase of the driving field naturally defines the $\hat{x}$ and $\hat{y}$ axes for the spin degree of freedom, our protocol can equivalently be viewed as preparing the collective Bloch vector at analogous positions on the pseudospin Bloch sphere.

Model and simulations

The dynamics of the experimental system are modelled by a master equation for the density operator $\hat{\rho}$ of the complete atom–light system,

$$\frac{d\hat{\rho}}{dt} = -\frac{i}{\hbar}[\hat{H}_{tot}, \hat{\rho}] + \mathcal{L}_c[\hat{\rho}] + \mathcal{L}_{el}[\hat{\rho}] + \mathcal{L}_s[\hat{\rho}] \tag{3}$$

Here, the Hamiltonian $\hat{H}_{tot} = \hat{H}_A + \hat{H}_L + \hat{H}_{AL}$ is split into three contributions characterizing the atoms, pumping of the cavity field and atom–light interaction respectively:

$$\hat{H}_A = \frac{\omega_a}{2}\sum_i \hat{\sigma}_i^z \tag{4}$$

$$\hat{H}_L = \omega_c \hat{a}^\dagger \hat{a} + \Omega_p(\hat{a}e^{i\omega_p t} + \hat{a}^\dagger e^{-i\omega_p t}) \tag{5}$$

$$\hat{H}_{AL} = \sum_i g_i(\hat{a}\hat{\sigma}_i^+ + \hat{a}^\dagger\hat{\sigma}_i^-) \tag{6}$$

where $\hat{a}$ ($\hat{a}^\dagger$) is the annihilation (creation) operator of the cavity mode and the sums are taken over $i = 1,\ldots,N$ atoms. To reiterate, ω_a is the frequency of the atomic transition, ω_c the frequency of the relevant cavity mode, Ω_p the effective amplitude of the injected field and ω_p the corresponding frequency. The spatial dependence of the coupling is characterized by $g_j = g\cos(kj)$ and $k = \pi\lambda_L/\lambda_c$, where $2g$ is the single-photon Rabi frequency at an antinode of the cavity mode. This form arises because the magic wavelength of the 1D optical lattice λ_L = 813 nm is

incommensurate with the wavelength $\lambda_c = 689$ nm of the cavity mode to which the atomic transition is coupled. For simplicity, we take the summation to run over $i = 1,2,\ldots,N$ total lattice sites, such that each site is assumed to be occupied by only a single atom. In reality, there are about 10^3 relevant lattice sites, and each is occupied by about 10^2–10^3 atoms, but as we assume contact interactions are not relevant and the atom–light coupling is consistent across the entire atomic sample, this simplification is reasonable.

Decoherence due to leakage of photons from the cavity at rate κ is described by the Lindblad term

$$\mathcal{L}_c[\hat{\rho}] = \frac{\kappa}{2}(2\hat{a}\hat{\rho}\hat{a}^\dagger - \hat{a}^\dagger\hat{a}\hat{\rho} - \hat{\rho}\hat{a}^\dagger\hat{a}) \tag{7}$$

while spontaneous emission on the atomic transition at rate γ and single-particle homogeneous broadening of the ensemble at rate γ_{el} are described by

$$\mathcal{L}_s[\hat{\rho}] = \frac{\gamma}{2}\sum_i 2\hat{\sigma}_i^-\hat{\rho}\hat{\sigma}_i^+ - \hat{\sigma}_i^+\hat{\sigma}_i^-\hat{\rho} - \hat{\rho}\hat{\sigma}_i^+\hat{\sigma}_i^- \tag{8}$$

$$\mathcal{L}_{el}[\hat{\rho}] = \frac{\gamma_{el}}{2}\sum_i \hat{\sigma}_i^z\hat{\rho}\hat{\sigma}_i^z - \hat{\rho} \tag{9}$$

The latter is attributed to a range of effects, including undesirable motion of the atoms in the optical lattice, and is discussed in more detail in the Supplementary Information.

The simulations presented in Figs. 2–4 are the result of numerical solution of equation (3) within the mean-field approximation (with the exception of the lower panels of Fig. 2a) which include additional effects due to axial motion that are discussed in the Supplementary Information. Specifically, we solve the equations of motion for $\boldsymbol{\sigma}_i \equiv (\langle\hat{\sigma}_i^x\rangle, \langle\hat{\sigma}_i^y\rangle, \langle\hat{\sigma}_i^z\rangle)$ and $\langle\hat{a}\rangle$, and factorize higher-order moments of the operators, for example, $\langle\hat{\sigma}_i^x\hat{\sigma}_j^y\rangle \equiv \langle\hat{\sigma}_i^x\rangle\langle\hat{\sigma}_j^y\rangle$. Further details regarding the numerical simulations can be found in the Supplementary Information.

The effective spin model that describes the nonlinear atomic dynamics throughout this manuscript is obtained from the atom–light model (equation (3)) by separate adiabatic elimination of the injected field and intracavity fluctuations, and the full calculation is detailed in the Supplementary Information. Here, we merely present the resulting Hamiltonian for the atoms:

$$\hat{H} = \hbar\sum_{i,j}\chi_{ij}\hat{\sigma}_i^+\hat{\sigma}_j^- + \hbar\sum_i\frac{\Omega_i}{2}\hat{\sigma}_i^x - \frac{\hbar\delta}{2}\sum_i\sigma_i^z \tag{10}$$

where $\chi_{ij} = -g_ig_j/\Delta$, $\Omega_i = -2g_i\Omega_p/\Delta$ with $\delta = \omega_p - \omega_a$ and $\Delta = \omega_c - \omega_a$. Moreover, we have assumed $|\Delta| \gg \kappa, g\sqrt{N}, \sqrt{g\Omega_p}, \delta$. In the limit $k = 2n\pi$ for $n \in \mathbb{Z}$, that is, uniform atom–light coupling $g_j \to g$, we recover the collective LMG model of equation (1).

Although in the experimental platform the atom–light coupling g_j is spatially varying owing to the incommensurate cavity and lattice wavelengths, the qualitative physics we explore is still consistent with the framework of the collective LMG model. Specifically, while the simulations of Figs. 2–4 take the proper form of g_j into account (see Supplementary Information), we observe that features of the detailed inhomogeneous model such as the critical point and dynamical timescales are consistent with the collective model upon a rescaling of the atom–light coupling.

For weak drives deep in the ferromagnetic phase, the collective model replicates the quantitative predictions of the inhomogeneous model upon replacement of the atom–light coupling with the r.m.s. average, $g \to g/\sqrt{2}$ and thus $\chi \to \chi/2$ and $\Omega \to \Omega/2$. This approximation is supported by comparison to experimental results for the period of the weak oscillations deep in the ferromagnetic phase, which are expected to be proportional to $1/(\chi N)$. In Extended Data Fig. 2a we extract this period from the experimental data as a function of cavity detuning Δ, which is equivalent to varying the interaction strength $\chi \propto 1/\Delta$. We confirm that the fitted slope agrees with the $\chi \to \chi/2$ correction for inhomogeneous atom–light coupling.

As the drive is increased, the rescaling required for quantitative comparison changes. Specifically, comparing to the critical point $\Omega_c^{theory}/(\chi N)$ obtained from a numerical calculation of the inhomogeneous model in the absence of decoherence, we find that the corresponding collective model requires a rescaling $g \to 0.62g$, and thus $\chi \to 0.38\chi$ and $\Omega \to 0.62\Omega$, to match the critical value $\Omega_c^{theory}/(\chi N) \approx 0.31$. The reduction of this value below the true collective critical drive $\Omega_c/(\chi N) = 1/2$ is consistent with that observed experimentally ($\Omega_c^{exp}/(\chi N) = 0.35(3)$).

Mapping the phase boundary

In Fig. 1b, c, we present the system phase diagram (under the assumption of uniform atom–light coupling), where we map the magnetization $\langle\hat{J}_z\rangle$ as a function of the probe detuning δ and drive amplitude Ω. A sharp boundary separates the dynamical phases for $\Omega/(\chi N) \lessapprox 0.65$, shown by the solid white line in Fig. 1c. However, as the drive is further increased and for $\delta/(\chi N) < -1/8$, the boundary becomes a smooth crossover, as shown by the dashed white line in Fig. 1c.

Using similar results to those shown in Fig. 3, for the inhomogeneous case relevant for experiment we are able to map out this boundary and define a critical detuning δ_c between the two dynamical phases for different fixed drive strengths Ω. We identify these values by looking at the maximum gradient on each of the experimental and numerical $\langle\hat{J}_z\rangle$ against δ plots shown in Fig. 3. In Extended Data Fig. 2b we plot δ_c against Ω (points) and compare with numerical simulations (solid lines) for two opposite cavity detunings $\Delta/(2\pi) = \pm 50$ MHz. For values above $\Omega/(\chi N) \approx 0.31$, the solid traces do not represent a strict phase boundary but rather characterize the crossover region, analogous to the crossover region in Fig. 1c for the homogeneous case.

In the Supplementary Information, we derive an expression for the boundary between the two dynamical phases based on the model presented in equation (1) in the mean field limit. In the homogeneous case, the phase boundary $\Omega_c(\delta)$ is, for $\delta/(\chi N) > -1/8$:

$$\frac{\Omega_c(\delta)}{\chi N} = \frac{1}{2}\left[2\left(1-\frac{\delta}{\chi N}\right)\left(1+\frac{2\delta}{\chi N}\right) - \frac{3}{2}\left(\frac{8\delta}{\chi N}+1\right) + \frac{1}{2}\left(1+\frac{8\delta}{\chi N}\right)^{3/2}\right]^{1/2} \tag{11}$$

To address the inhomogeneous coupling present in our experiment, we rescale $g \to 0.62g$ and thus $\chi \to 0.38\chi$ and $\Omega \to 0.62\Omega$ in this equation as described earlier. A comparison of the rescaled equation (11) to the experimental data is shown as the black traces in Extended Data Fig. 2b for two different detunings.

Data availability

Data relevant to the figures and conclusions of this manuscript are available at https://doi.org/10.5061/dryad.mgqnk98w9[51].

Code availability

The codes used in the analysis of experimental data and to carry out associated theoretical calculations are available from the corresponding authors upon reasonable request.

48. Norcia, M. A. & Thompson, J. K. Strong coupling on a forbidden transition in strontium and nondestructive atom counting. *Phys. Rev. A* **93**, 023804 (2016).
49. Norcia, M. A. et al. Frequency measurements of superradiance from the strontium clock transition. *Phys. Rev. X* **8**, 021036 (2018).
50. Norcia, M. A., Winchester, M. N., Cline, J. R. K. & Thompson, J. K. Superradiance on the millihertz linewidth strontium clock transition. *Sci. Adv.* **2**, e1601231 (2016).
51. Muniz Silva, J. A. et al. Exploring dynamical phase transitions with a cavity-QED platform, v2. Dryad dataset (2020); https://doi.org/10.5061/dryad.mgqnk98w9.

Acknowledgements We acknowledge discussions with I. Spielman, M. Holland and A. Shankar. This work is supported by the Air Force Office of Scientific Research (AFOSR) grant FA9550-18-1-0319, by the Defense Advanced Research Projects Agency (DARPA) Extreme Sensing and ARO grant W911NF-16-1-0576, the ARO single investigator award W911NF-19-1-0210, the US National Science Foundation (NSF) PHY1820885, NSF JILA-PFC PHY-1734006 grants, and by the National Institute of Standards and Technology (NIST). J.R.K.C. acknowledges financial support from NSF GRFP.

Author contributions J.A.M., D.J.Y., J.R.K.C. and J.K.T. collected and analysed the experimental data. R.J.L.-S., D.B. and A.M.R. developed the theoretical model. All authors discussed the results and contributed to the preparation of the manuscript.

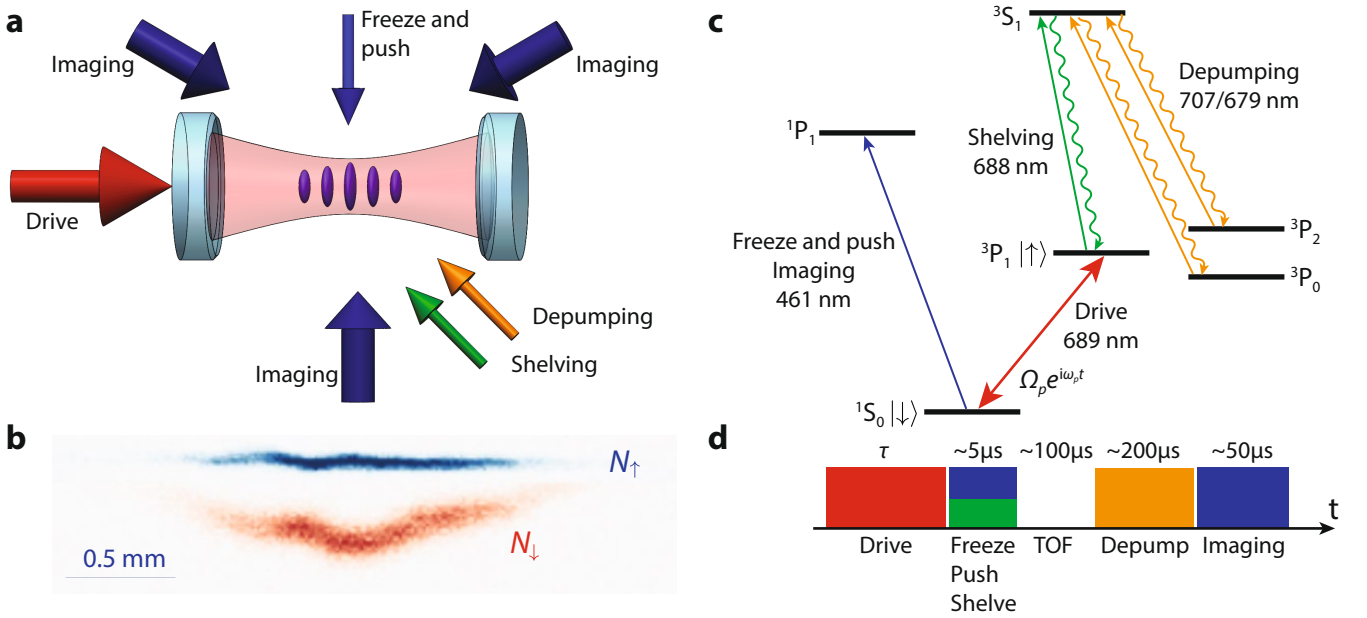
Competing interests The authors declare no competing interests.

Additional information

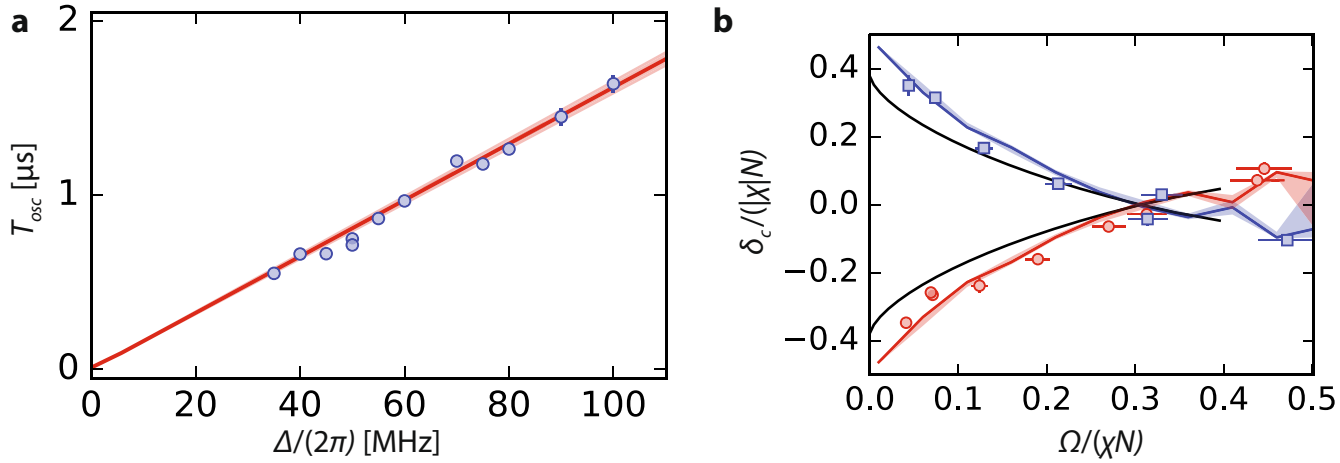
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2224-x.

Correspondence and requests for materials should be addressed to A.M.R. or J.K.T.

Peer review information *Nature* thanks Murray Barrett, Maria Luisa Chiofalo and Farokh Mivehvar for their contribution to the peer review of this work.



Extended Data Fig. 1 | Experimental platform. a, An optical cavity is driven by a 689-nm coherent field that establishes an intra-cavity field $\Omega_p e^{i\omega_p t}$, which is near resonance with the $^1S_0$ to $^3P_1$ transition in ^{88}Sr. Inside the cavity, an ensemble of atoms is confined in a 1D optical lattice at 813 nm. Different lasers are applied for shelving excited-state atoms into long-lived metastable excited states, for freezing the system dynamics, for applying a radiation pressure force that pushes ground states in a direction transverse to the cavity axis, for optically pumping atoms from long lived metastable excited states back to the ground state, and for fluorescence imaging of atoms in the ground state. **b**, A typical fluorescence image captured on a CCD, showing the state-resolved imaging technique. The N_e excited state atoms that were shelved into $^3P_{0,2}$ while the freeze/push beam was applied remain near the trapping region. The N_g ground-state atoms are pushed away from the trapping region. Based on their spatial location, the atoms assigned to be in the excited (ground) state are shown in false colour blue (orange). **c**, The relevant energy levels for ^{88}Sr, the laser wavelengths and their functions. **d**, Experimental timing sequence and typical timescales.



Extended Data Fig. 2 | Probing many-body dynamics and mapping the phase boundary. a, Oscillation period as function of the cavity detuning Δ for $2\Omega_p/(Ng) = 0.104(4)$, $\delta = 0$ and atoms starting in $|\downarrow\rangle$. Blue points are experimental values, solid red line represents the mean-field prediction for the same drive and atom number, and the shaded red area represents typical experimental fluctuations on $2\Omega_p/(Ng)$. The period is extracted from sinusoidal fits to data as in Fig. 2a, after removing a linear term caused by the single-particle dephasing effects. The mean-field value (red solid line) is $T_{osc} = 2\pi/(N\chi)$ with the effective replacements due to inhomogeneous coupling as discussed in Methods. Measurements are taken in the dispersive limit where $\Delta \gg \sqrt{N}g$. **b**, Critical detuning δ_c as function of the drive Δ for $\Delta/(2\pi) = \pm 50$ MHz (red and blue points, respectively). We also plot the theoretical prediction for the phase boundary (equation (11)) with rescaled parameters, and predictions of the numerical model (solid lines) including uncertainty based on the typical fluctuations in $\Omega/(\chi N)$. Error bars are statistical (1σ).

Article

Electrical manipulation of a topological antiferromagnetic state

https://doi.org/10.1038/s41586-020-2211-2

Received: 17 November 2019

Accepted: 8 March 2020

Published online: 20 April 2020

Check for updates

Hanshen Tsai[1,2,8], Tomoya Higo[1,2,8], Kouta Kondou[2,3], Takuya Nomoto[2,4], Akito Sakai[1,2], Ayuko Kobayashi[1], Takafumi Nakano[2,5], Kay Yakushiji[2,5], Ryotaro Arita[2,3,4], Shinji Miwa[1,2,6], Yoshichika Otani[1,2,3,6] & Satoru Nakatsuji[1,2,6,7]✉

Electrical manipulation of phenomena generated by nontrivial band topology is essential for the development of next-generation technology using topological protection. A Weyl semimetal is a three-dimensional gapless system that hosts Weyl fermions as low-energy quasiparticles[1–4]. It has various exotic properties, such as a large anomalous Hall effect (AHE) and chiral anomaly, which are robust owing to the topologically protected Weyl nodes[1–16]. To manipulate such phenomena, a magnetic version of Weyl semimetals would be useful for controlling the locations of Weyl nodes in the Brillouin zone. Moreover, electrical manipulation of antiferromagnetic Weyl metals would facilitate the use of antiferromagnetic spintronics to realize high-density devices with ultrafast operation[17,18]. However, electrical control of a Weyl metal has not yet been reported. Here we demonstrate the electrical switching of a topological antiferromagnetic state and its detection by the AHE at room temperature in a polycrystalline thin film[19] of the antiferromagnetic Weyl metal Mn_3Sn[9,10,12,20], which exhibits zero-field AHE. Using bilayer devices composed of Mn_3Sn and nonmagnetic metals, we find that an electrical current density of about 10^{10} to 10^{11} amperes per square metre induces magnetic switching in the nonmagnetic metals, with a large change in Hall voltage. In addition, the current polarity along the bias field and the sign of the spin Hall angle of the nonmagnetic metals—positive for Pt (ref. [21]), close to 0 for Cu and negative for W (ref. [22])—determines the sign of the Hall voltage. Notably, the electrical switching in the antiferromagnet is achieved with the same protocol as that used for ferromagnetic metals[23,24]. Our results may lead to further scientific and technological advances in topological magnetism and antiferromagnetic spintronics.

Recent extensive studies in condensed matter physics have revealed various novel quantum phases with nontrivial topology in the electronic band structure[4,25,26]. One example of such topological systems is a Weyl semimetal[1–4]. Two non-degenerate bands touch linearly at a pair of momentum points, forming gapless Weyl fermions with different chiralities in a state that breaks time-reversal symmetry (TRS) or inversion symmetry. These touching points, or Weyl nodes, act as topologically protected, unit-strength (anti)monopoles of underlying Berry curvature, and lead to various emergent phenomena, such as large AHE, anomalous Nernst effect (ANE), chiral anomaly and optical gyrotropy[1–16].

To develop science and technology using topological states, a crucial next step would be to manipulate these emergent phenomena electrically. In a Weyl semimetal, such manipulation can be made by moving the Weyl points around in the Brillouin zone. TRS-breaking or magnetic Weyl semimetals would be suitable for this purpose owing to their magnetic texture. Besides, antiferromagnets (AFMs) have recently attracted considerable attention as the active material for next-generation spintronics devices, with the prospect of providing higher density storage and much faster operation speed than their ferromagnetic counterparts[17,18]. However, there have been no reports on electrical manipulation of either antiferromagnetic or ferromagnetic Weyl semimetals.

The advancement in our understanding of topological aspects in the electronic structure has led to the discovery of the AHE[27–29] in non-ferromagnetic systems such as spin liquids[30] and chiral AFMs Mn_3X (X = Sn, Ge, Ga, Ir, Pt, Rh)[9,31–35]. This discovery has shown that AFMs may exhibit large transverse responses, such as the AHE and ANE, induced by the Berry curvature in momentum space in the absence of magnetization M (refs. [9–11,31–35]). In particular, theoretical studies and experiments using single crystals have demonstrated that Mn_3Sn hosts magnetic Weyl fermions[9,10,12,20]. Large topological responses such as the AHE and the ANE, which are associated with the topological protection of Weyl nodes, are robust against disorder, impurities and thermal fluctuation; for example, they appear over a wide range of Mn concentrations[9,10,12] and even in polycrystalline thin films[19,36], paving the way for future applications.

[1]Institute for Solid State Physics, University of Tokyo, Kashiwa, Japan. [2]CREST, Japan Science and Technology Agency, Kawaguchi, Japan. [3]Center for Emergent Matter Science (CEMS), RIKEN, Wako, Japan. [4]Department of Applied Physics, University of Tokyo, Tokyo, Japan. [5]Spintronics Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. [6]Trans-scale Quantum Science Institute, University of Tokyo, Tokyo, Japan. [7]Department of Physics, University of Tokyo, Tokyo, Japan. [8]These authors contributed equally: Hanshen Tsai, Tomoya Higo. ✉e-mail: satoru@phys.s.u-tokyo.ac.jp

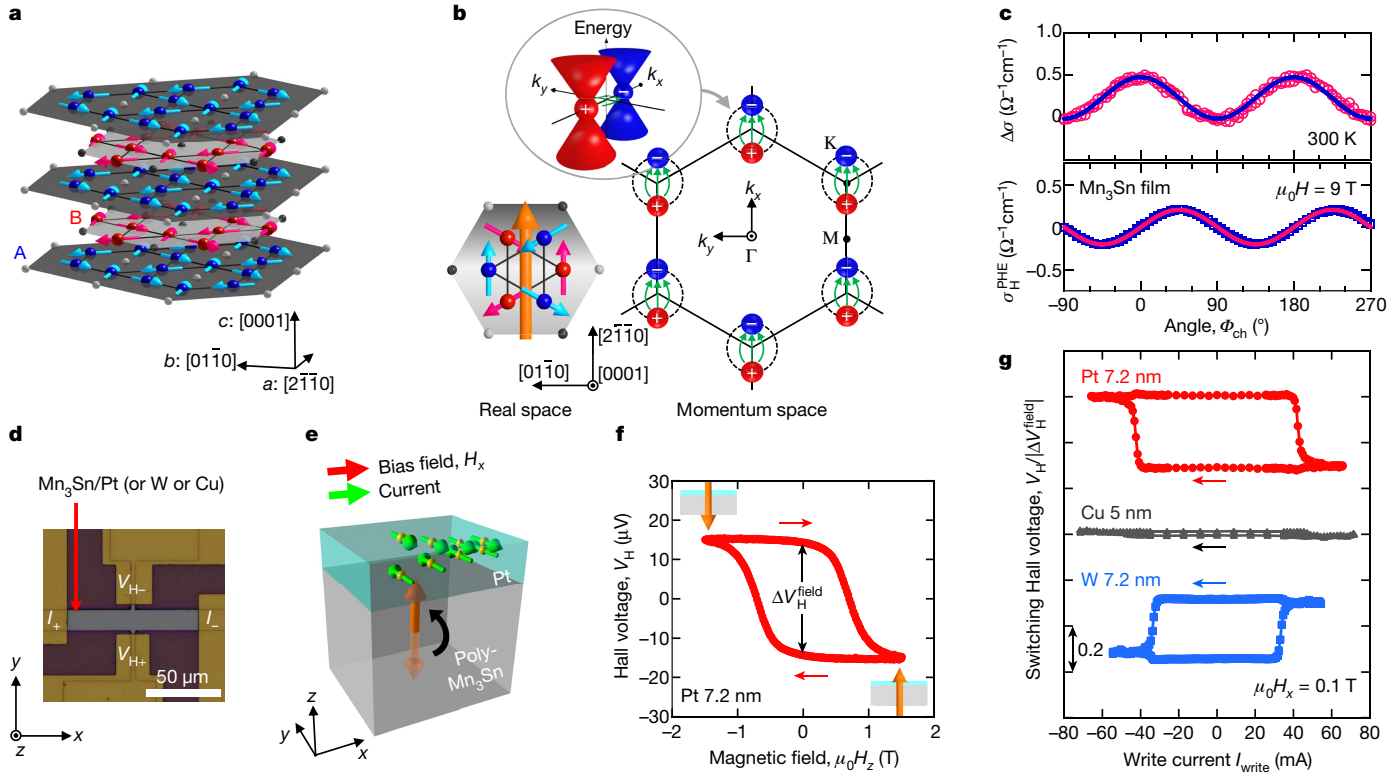


Fig. 1 | Topological Weyl AFM Mn_3Sn and bilayer device layout. a, Mn_3Sn crystal structure and inverse triangular spin (ITS) structure. The large blue and red spheres (small grey and black spheres) represent Mn (Sn) atoms at $z = 0$ and 1/2, respectively. The Mn magnetic moments (light blue and pink arrows) lie within the kagome layer with an ABAB stacking sequence and form the ITS structure at room temperature. The spin structure on the kagome bilayers can be viewed as ferroic ordering of a cluster magnetic octupole. **b**, Left, cluster magnetic octupole (orange arrow) consisting of six spins on the kagome bilayer. Right, schematic distribution of the Weyl points near the Fermi energy in momentum space (k_x–k_y plane at $k_z = 0$) for the magnetic structure shown in the left illustration. Red and blue spheres correspond to Weyl nodes that act as sources (+) and drains (−), respectively, of the Berry curvature (green arrows)[12]. Inset, three-dimensional schematic of a pair of Weyl nodes. **c**, Angular dependence of the longitudinal magnetoconductivity $\Delta\sigma = \sigma - \sigma_\perp$ and the planar Hall conductivity σ_H^{PHE} of the Mn_3Sn film at 300 K and 9 T. Blue and pink solid lines for $\Delta\sigma$ and σ_H^{PHE} are fitting results using theoretical equations for the chiral anomaly, that is, $\Delta\sigma = \Delta\sigma_{chiral}\cos^2\Phi_{ch}$ and $\sigma_H^{PHE} = \Delta\sigma_{chiral}\sin\Phi_{ch}\cos\Phi_{ch}$, respectively[13,16]. Here $\Delta\sigma_{chiral}$ ($\Delta\sigma_{chiral} = \sigma_\parallel - \sigma_\perp$) is the positive magnetoconductivity induced by the chiral anomaly; $\sigma_\parallel$ and $\sigma_\perp$ is the magnetoconductivity when the current is parallel ($\Phi_{ch} = 0°$) and perpendicular ($\Phi_{ch} = 90°$) to the magnetic field, respectively; Φ_{ch} refers to the in-plane rotation angle between the magnetic field and the electrical current directions, as shown in Extended Data Fig. 2c. See Methods for details. **d**, Optical micrograph of the fabricated Mn_3Sn NM bilayer Hall bar devices. Write and read currents are applied along the x direction under magnetic field H_x, H_y or H_z along the x, y or z direction, respectively. **e**, SOT switching. The spin-polarized current (green arrows on yellow spheres) generated in Pt exerts an SOT and causes switching of the polarization axis of the cluster magnetic octupole (orange arrow) in the polycrystalline Mn_3Sn under a write current and a bias field along the x direction. **f**, Hall voltage V_H versus magnetic field along the z direction, H_z, for the Mn_3Sn/Pt (7.2 nm) device at room temperature. The insets illustrate the direction of the cluster magnetic octupole. **g**, Hall voltage V_H versus write current I_{write} for the Mn_3Sn/Pt (7.2 nm), Mn_3Sn/Cu (5 nm) and Mn_3Sn/W (7.2 nm) devices at room temperature. The Hall voltage is normalized by the zero-field Hall voltage ΔV_H^{field} obtained from the magnetic field dependence for each sample.

Here we demonstrate the electrical switching of the topological antiferromagnetic state in a polycrystalline thin film of Mn_3Sn, which we detect through the AHE. Our work is important not only because it provides successful electrical manipulation of the AHE in a topological Weyl metal but also because it applies the same switching protocol as that used for conventional spintronics[23,24]. Mn_3Sn is one of the best-studied kagome-based metals, which have attracted substantial interest owing to their nontrivial band topology[9,15,37] and unusual magnetic responses at interfaces[38–40]. Besides, the effect of spin injection into Mn_3Sn has been theoretically predicted, promoting its application in spintronics[39,40].

Mn_3Sn has the hexagonal structure DO_{19} with ABAB stacking of a (0001) kagome layer of Mn, and the geometrical frustration leads to a three-sublattice non-collinear antiferromagnetic ordering of Mn spins below the Néel temperature of $T_N \approx 430$ K (ref. [9]; Fig. 1a). This antiferromagnetic spin texture can be viewed as a ferroic ordering of a cluster magnetic octupole T (Fig. 1b, left)[41], which gives rise to M-linear responses, such as AHE[9], ANE[10] and the magneto-optical Kerr effect[42], instead of a very small uncompensated magnetization of $M \approx 0.006\mu_B$ per formula unit (f.u.; μ_B, Bohr magneton) induced by the spin canting within the (0001) plane. The strikingly large AHE and ANE in Mn_3Sn—which are comparable with, or even exceed, those in ferromagnets (FMs)—come from the Weyl fermions and the large Berry curvature, which is equivalent to a few hundreds of teslas[10,12,20]. Moreover, the polarization direction of the magnetic octupole determines the location of Weyl nodes and the associated distribution of the Berry curvature in momentum space (Fig. 1b, right)[12]. Thus, it is essential to control the orientation of the magnetic octupole to achieve electrical manipulation of topological responses.

Before discussing the electrical switching measurements, we first demonstrate that our thin films host the same topological Weyl semimetal state as in the Mn_3Sn single crystals[10,12,20]. For this, we employ two types of transport probes for the Weyl fermions: the chiral anomaly and the ANE[1,4,6,8,10–14,16] (Methods). First, our measurements using thin films find magneto-transport phenomena that are fully consistent with the chiral anomaly caused by the Weyl fermions (Extended Data

Figs. 1, 2). For instance, the angular dependence of the magnetoconductivity σ and the planar Hall conductivity σ_H^{PHE} are well fitted by the theoretical equations for the chiral anomaly[13,16] (Fig. 1c). Second, the ANE probes the Berry curvature at the Fermi energy E_F (ref. [29]). Thus, the Weyl points near E_F could enhance the ANE much more than what would be expected on the basis of the empirical scaling law with magnetization typically used for FMs[10,11,14]. Our measurements using the Mn_3Sn thin films reveal that this is indeed the case (Extended Data Figs. 3, 4a), confirming the magnetic Weyl semimetal state realized in the thin films. Besides, we find that the associated transverse thermoelectric conductivity α_N changes its sign by reversing the magnetic field, following the same field dependence as the Hall conductivity (Extended Data Fig. 4b). This further demonstrates that, similarly to α_N, the switching of the AHE effectively detects the rotation of the pairs of Weyl points and the associated Berry curvature in momentum space (Extended Data Fig. 5).

To generate a spin current by the spin Hall effect (SHE), we employ nonmagnetic metals (NMs; NM = Pt, W, Cu) in a device with the structure $Si/SiO_2/Ru(2)/Mn_3Sn(40)/NM(d_{NM})/AlO_x(5)$ (numbers in parentheses denote thickness in nanometres) deposited on a Si/SiO_2 substrate (Methods, Fig. 1d, e). First, we measure the Hall voltage V_H as a function of the out-of-plane magnetic field H_z to quantitatively estimate the population of switchable domains normal to the plane of the Mn_3Sn layer in the device. Figure 1f shows the clear hysteresis of the Hall voltage with the zero-field change of ΔV_H^{field} ($\Delta V_H^{field} = V_H(+H_z \to 0) - V_H(-H_z \to 0)$; ref. [19]). As the polarization direction of the octupole points to the same direction as the tiny canted moment, the results indicate that the negative (positive) value of V_H is generated by the '$+z$ ($-z$) domain' with the positive $+z$ (negative $-z$) component of the polarization direction of the octupole.

Now we examine possible electrical switching for the bilayer device. A 100-ms write pulse current I_{write} followed by a d.c. read current of $I_{read} = 0.2$ mA is applied along the x direction (Fig. 1d, Extended Data Fig. 6a). Here, the current I flows in the entire stack, and the current density J corresponds to the part flowing in the NM layer. Remarkably, the electrical current that flows in the device results in different signs of electrical switching according to the sign of the spin Hall angle θ_{SH} for the NM layer (Fig. 1g). As for NM = Pt ($\theta_{SH} > 0$)[21], a clear negative (positive) jump appears in the Hall voltage under a positive (negative) current larger than a critical threshold of the write current I_c in a bias field H_x along direction x. The magnitude of the jump reaches approximately 30% of the total Hall voltage change $|\Delta V_H^{field}|$ in the field sweep measurements. As W has a large θ_{SH} with an opposite sign ($\theta_{SH} < 0$) to that of Pt (ref. [22]), we carry out the switching experiments by replacing Pt by W (Fig. 1d). Importantly, the W device has a switching polarity opposite to that of the Pt device, and the magnitude of the voltage jump amounts to around 25% of $|\Delta V_H^{field}|$ (Fig. 1g). We have also fabricated a device with Cu (Fig. 1g) and a device without the NM layer with the same configuration, and found almost no hysteresis with the electrical current cycle (Extended Data Fig. 7a–c). The difference in switching polarity between the Pt and W devices and the absence of switching in the Cu and no-NM-layer devices cannot be accounted for by the Oersted field generated by the electrical current and the spin–orbit torque (SOT) due to the SHE in the Mn_3Sn layer, but agree well with the sign of θ_{SH} in the NM layer. These results demonstrate that the SOT from the SHE of the NM layer induces the perpendicular switching of the antiferromagnetic domain.

Generally, the sign of the SOT switching is determined by the bias field along the I_{write} direction. To examine this under various directions of the bias field, we measure the Hall voltage V_H as a function of I_{write} (V_H–I_{write} loop). Figure 2a clearly shows that if the directions of I_{write} and H_x are the same, the voltage exhibits a negative jump at $|I_c|$, accompanied by an increase in the population of $+z$ domains with positive z-direction component of the cluster magnetic octupole. If I_{write} and H_x have opposite directions, the jump becomes positive, increasing the population of $-z$ domains. In addition, the magnitude of the jump, $\Delta V_H^{current} = V_H(+I_{write} \to 0) - V_H(-I_{write} \to 0)$, changes with the bias field along the current direction (x; Fig. 2b, c). The same switching measurements performed under bias magnetic fields parallel to y and z indicate that only the bias field along x induces switching of the antiferromagnetic domains. This observation follows the expectation of the symmetry requirement of the SOT switching of the perpendicular magnetization[23,24]. The NM thickness dependence of the switching Hall voltage $|\Delta V_H^{current}|$ in the Pt and W devices shows a systematic change over 20 devices with a saturation of around 30% of $|\Delta V_H^{field}|$ at $d_{NM} \geq 2$ nm, indicating robust switching properties over a wide range of NM thicknesses (Fig. 2d, Extended Data Fig. 6b).

The switching or critical write current density J_c is found to be reasonably small. It is estimated to be 2×10^{11} A m^{-2} for 7.2-nm-thick Pt and 5×10^{10} A m^{-2} for 7.2-nm-thick W devices. These are smaller than the values originally reported for the first observations of electrical switching in NM/FM devices ($J_c \approx 10^{12}$ A m^{-2})[23,24] and comparable to recent estimates for AFM/FM devices (with the AFM layer as a spin current source)[43], collinear Néel–SOT devices[44,45] and collinear AFM/Pt devices[46,47]. The estimated heating of the central part of the device is approximately 50 K, and the temperature remains lower than the Néel point even when the write current I_{write} is on. Moreover, V_H is not affected by heating due to I_{write} injection because of the sufficiently long wait time of 600 ms (Extended Data Figs. 6a, 7d, Methods). We carry out the same switching measurements at temperatures lower than room temperature (200, 250 and 295 K; Fig. 2e). Notably, the same switching takes place and the threshold current increases only slightly. All our experiments confirm that our bilayer devices exhibit deterministic magnetic switching due to the SOT exerted on the antiferromagnetic spin texture. This switching can also support reproducible bipolar writing as an antiferromagnetic memory. Alternate pulse currents with different signs systematically produce magnetic switching of the reading V_H over 200 times (Fig. 3a) and the signal is stable against consecutive injections of $I_{write} > I_c$ (Extended Data Fig. 7e), demonstrating its controllability.

To understand the deterministic switching mechanism based on the SOT, we study the dynamics of the sublattice moments $\mathbf{m}_{ia}$ on the one-layer kagome lattice (blue layer on the y–z plane in Fig. 4a) obeying the Landau–Lifshitz–Gilbert (LLG) equation[21,48]

$$\dot{\mathbf{m}}_{ia} = -|\gamma|\mathbf{m}_{ia} \times \mathbf{H}_{\mathrm{eff},ia} + \alpha \mathbf{m}_{ia} \times \dot{\mathbf{m}}_{ia} + \mathbf{T}_{ia} \quad (1)$$

where the suffix i denotes a unit cell and $a = (1, 2, 3)$ a sublattice. Here, the effective magnetic field $\mathbf{H}_{\mathrm{eff},ia}$ is defined as $\mathbf{H}_{\mathrm{eff},ia} = -M_S^{-1}\delta\mathcal{H}/\delta\mathbf{m}_{ia}$ with $M_S = 3\mu_B$ the saturation magnetic moment of a Mn atom and $\mathcal{H}$ the Hamiltonian (see Methods). The first and second terms of the right-hand side of equation (1) represent the gyroscopic torque and the Gilbert damping torque, respectively, γ ($\gamma < 0$) is the gyromagnetic ratio of the electron and α denotes the Gilbert damping coefficient. The third term represents the external torque due to the spin injection, namely, the spin-transfer-induced in-plane torque, $\mathbf{T}_{ia} = \frac{\hbar|\gamma|J_0\theta_{SH}}{2em_S d}\mathbf{m}_{ia} \times (\mathbf{y} \times \mathbf{m}_{ia})$ caused by the SOT. Here, θ_{SH} denotes the spin Hall angle in Pt (ref. [21]), d is the thickness of the Mn_3Sn layer, $m_S = 6M_S$ per unit cell volume, J_0 is the current density and $\mathbf{y} = (0, 1, 0)$ is the unit vector along the y axis.

Given the polycrystalline character of our Mn_3Sn layer, there are three crystal grain configurations representing the Mn_3Sn Hall bar devices (see Methods section 'Crystal grain configurations of polycrystalline Mn_3Sn Hall bar devices', Extended Data Figs. 8, 9). Here, we focus on configuration (a), with the kagome layer perpendicular to I and parallel to the polarization direction $\mathbf{p}$ (Fig. 4a); $\mathbf{p}$ is anti-parallel to the direction of the spin angular momentum of spin accumulation at the Mn_3Sn/Pt interface, because we find that deterministic switching accompanied by the AHE change may arise only in this configuration (Methods). For the simulation of the magnetization switching process, we set parameters consistent with the experimentally obtained physical parameters in

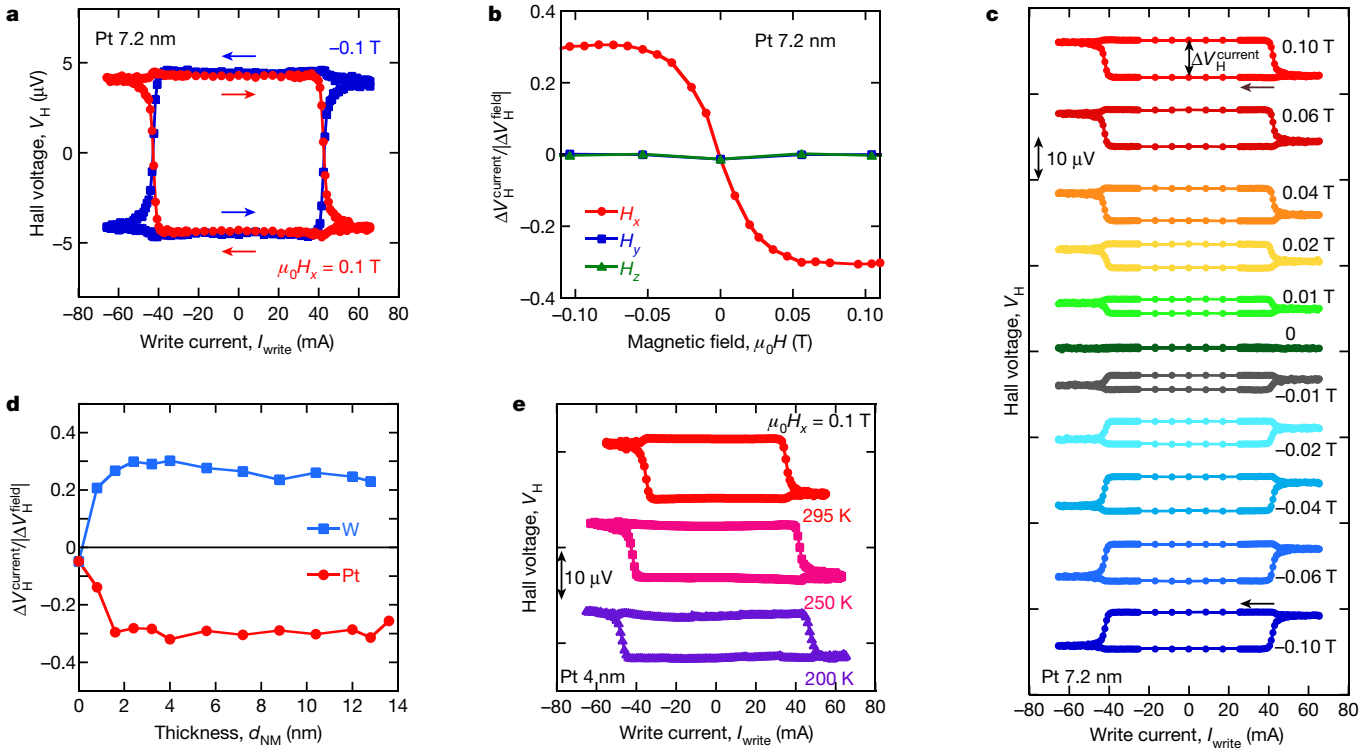


Fig. 2 | SOT-induced magnetic switching in the Mn_3Sn devices. **a**, Dependence of the Hall voltage V_H on the write current I_{write} (V_H–I_{write} loops) for the Mn_3Sn/Pt (7.2 nm) device under a magnetic field of H_x = +0.1 T (red) and −0.1 T (blue) along the x direction. **b**, Magnetic field dependence of the ratio of the current-induced Hall voltage switching to the field-induced Hall voltage switching, $\Delta V_H^{current}/|\Delta V_H^{field}|$, under a field applied along the x, y and z directions for the Mn_3Sn/Pt (7.2 nm) device. **c**, V_H–I_{write} loops for the Mn_3Sn/Pt (7.2 nm) device under various magnetic fields H_x. **d**, Dependence of $\Delta V_H^{current}/|\Delta V_H^{field}|$ on the thickness d_{NM} of the NM layer under H_x = 0.1 T. Measurements in **a**–**d** were performed at room temperature. **e**, V_H–I_{write} loops for the Mn_3Sn/Pt (4 nm) device under H_x = 0.1 T at 200, 250 and 295 K.

Mn_3Sn, where the sixfold magnetocrystalline anisotropy energy and the spontaneous magnetization are 310 J m^{-3} and 0.01μ_B per f.u., respectively (Methods). Figure 4b shows an example of numerical results for the evolution of the in-plane angle φ of the octupole polarization T as a function of time t for the case $\mathbf{p}\|y$ and $\mathbf{H}\|x$. Here, φ refers to the angle within the kagome plane (y–z plane in Fig. 4a) measured from the y direction (φ = 0) parallel to $\mathbf{p}$ (Fig. 4c, Extended Data Fig. 9a). In Fig. 4b, we show the results calculated with an initial spin structure in which the octupole polarization points to $\varphi = \pi/6$, one of the easy-axis directions.

We illustrate the switching dynamics schematically in Fig. 4d–f. Before turning on the write current I_{write}, we set the initial spin state to have $\varphi = \pm\pi/6$, as shown in Fig. 4d. Subsequently, the application of I_{write} exerts in-plane spin-transfer torques $\mathbf{T}_{ia}$. As a consequence, torque of the form $\mathbf{T}'_{ia} = \alpha\mathbf{m}_{ia} \times \mathbf{T}_{ia}$ is exerted on the sublattice moments $\mathbf{m}_{ia}$ in the direction of $\mathbf{m}_{ia} \times \mathbf{p}$, which thus acquire out-of-plane components. Unlike the SOT switching of ferromagnetic magnetization, where $\mathbf{T}_{ia}$ dominates, the rotational motion of the octupole polarization T is determined by the sum of the precessional movements of the sublattice

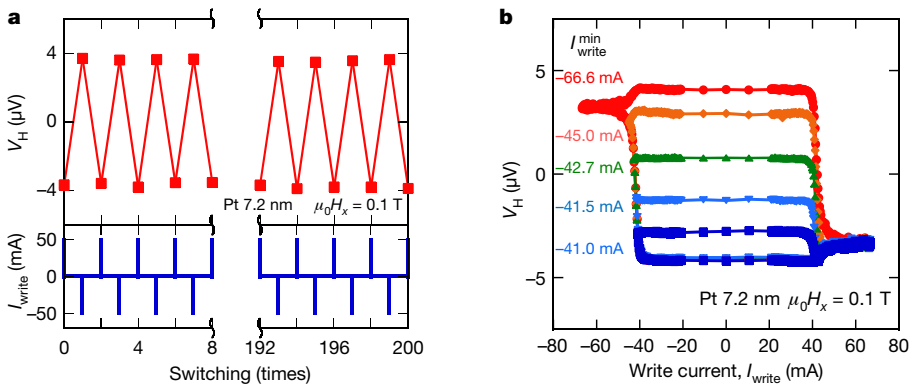


Fig. 3 | Reconfigurable antiferromagnetic switching. **a**, Hall voltage V_H (top) and applied write current I_{write} (bottom) for the Mn_3Sn/Pt (7.2 nm) Hall bar devices at room temperature and under a magnetic field of H_x = 0.1 T along the x direction for 200 times of switching between positive and negative I_{write} with a duration of 100 ms. The Hall voltage V_H measured at I_{read} = 0.2 mA changes sign depending on the polarity of I_{write} = 50 mA. **b**, V_H–I_{write} loops for the Mn_3Sn/Pt (7.2 nm) device under H_x = 0.1 T at room temperature. The minimum write current I^{min}_{write} affects the magnitude of the current-induced Hall voltage switching.

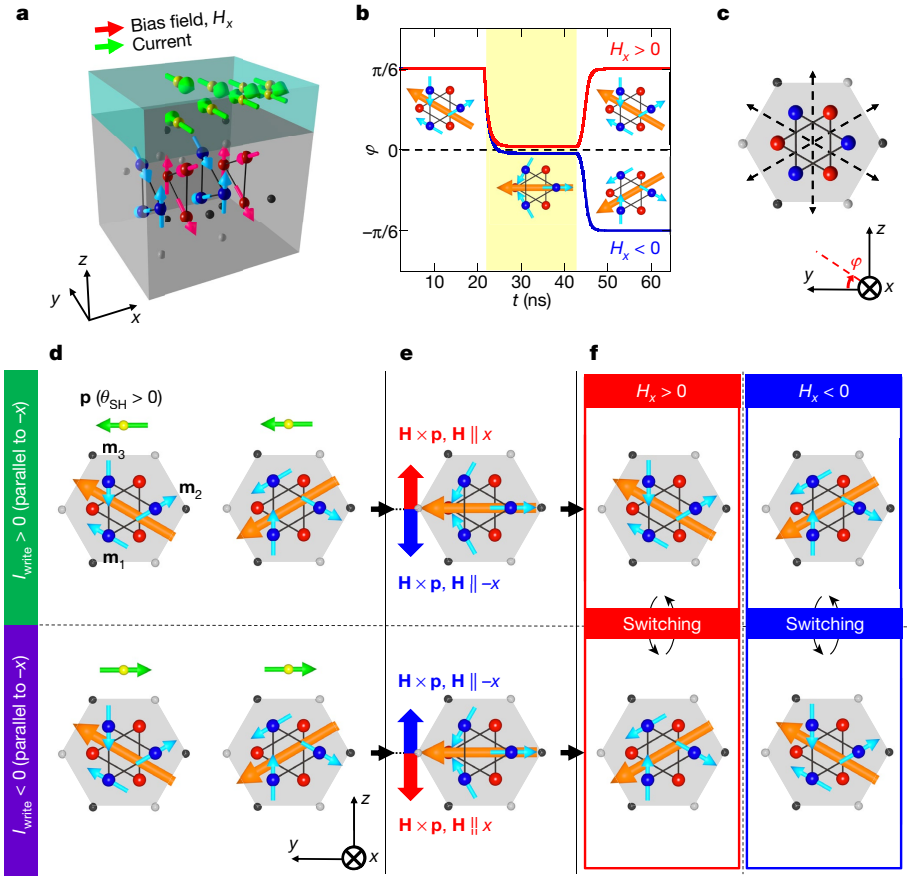


Fig. 4 | SOT mechanism and electrical switching of noncollinear spin texture. a, SOT switching in the crystal grain configuration (a) (see Methods section 'Crystal grain configurations of polycrystalline Mn_3Sn Hall bar devices', Extended Data Figs. 8, 9) under a current and a bias magnetic field **H** along the *x* direction.The *c* plane (the kagome layer) (0001) is perpendicular to the current and the *b* direction [01$\bar{1}$0] is parallel to **p**. **p** is parallel to the spin moments of the spin current (green arrows, antiparallel to the spin angular momentum). **b**, In-plane angle φ of the octupole polarization as a function of time *t* for **p** ∥ *y* and **H** ∥ *x*, calculated numerically using the LLG equation. A pulse current (damping-like torque $\mathbf{T}_{ia}$) is applied for 21.5 ns < *t* < 42.9 ns (yellow shaded region) and leads to the steady state with an angle of $\varphi = 0° \pm \delta$, the sign of which is determined by the sign of the bias field H_x. **c**, Crystal grains of Mn_3Sn in configuration (a). Dashed arrows represent the magnetic easy axis. **d**–**f**, Switching dynamics of the sublattice moments (light blue arrows) and octupole polarization (orange arrows) in Mn_3Sn with configuration (a) shown in **a**. **d**, The upper (lower) panel corresponds to the initial states, before the application of the write current I_{write} along the +*x* (−*x*) direction induces spin currents with polarization **p** (green arrow). **e**, Steady states under I_{write}. The switching direction of the octupole polarization is determined by **H** × **p** (blue and red arrows). **f**, Final states, determined by the sign of **H** (left, $H_x > 0$; right $H_x < 0$) and the sign of I_{write} (top, $I_{write} > 0$; bottom, $I_{write} > 0$). The black arrows indicate the switching induced by the I_{write} sweep under a positive (left) or negative (right) bias field.

moments $\mathbf{m}_{ia}$ driven by $\mathbf{T}'_{ia}$. The spin torque leads to the steady state under I_{write} by rotating the octupole polarization to $\varphi = 0 \pm \delta$ during the application of the write current I_{write}. In the absence of the bias field, the two initial states converge to the same steady state irrespective of the sign of I_{write} (Fig. 4d–e, Extended Data Fig. 10a, Methods).

Upon turning off I_{write}, the octupole polarization direction is relaxed and finally aligned along the closest easy axes (*a* axes) parallel to {2$\bar{1}\bar{1}$0}, selected by the sign of $\pm\delta$ and thus by the bias field direction. More precisely, the sign of $\pm\delta$ is determined by the torque exerted on the *c*-axis components of the sublattice moments $\mathbf{m}_\perp$ induced by the bias field H_x ($\mathbf{m}_\perp \times \mathbf{p} \parallel \mathbf{H} \times \mathbf{p}$) (Fig. 4e). As a result, the I_{write} sweep induces switching of the octupole polarization *T* between the *b* axes parallel to {01$\bar{1}$0} inclined by ±π/6 from the horizontal direction to the interface (Fig. 4f). This is analogous to the ferromagnetic case; using a perpendicularly magnetized ferromagnetic film, the SOT may cause deterministic switching between the two closest stable directions with the lowest anisotropy energy, that is, perpendicular directions. In our model, this corresponds to $\varphi = \pm\pi/6$ along the easy axes parallel to {2$\bar{1}\bar{1}$0} (Fig. 4f), and we have confirmed that this is always the case with realistic numerical conditions. As the steady state under I_{write} does not depend on the polarity of I_{write} and the final state is determined by the torque $\sim \boldsymbol{m}_\perp \times \mathbf{p} \parallel \mathbf{H} \times \mathbf{p}$, the same deterministic switching takes place between another pair of the easy axes, $\varphi = \pm 5\pi/6$.

If we focus on the initial and final states alone, the deterministic switching of the octupole polarization *T* in Mn_3Sn is generally analogous to that of the magnetization in FMs because the same symmetry requirements apply to both cases. However, we note that the emergence of the same steady states irrespective of the sign of I_{write} (Fig. 4e) is unique to our case because the switching dynamics is different from magnetization switching in FMs, where the steady states depend on the sign of I_{write} through the anti-damping SOT, $\mathbf{T}_{ia}$ (Methods, Extended Data Fig. 10b). Besides, our results indicate that the switching leads to a change in the AHE by $\sin(\pi/6) = 50\%$ of the total size expected for the field sweep measurement $|\Delta V_H^{field}|$, setting the maximum change in the AHE signal.

Importantly, the multi-grain character of Mn_3Sn allows us to tune the magnitude of the Hall voltage change using the write current I_{write} in an analogue manner. Figure 3b shows the V_H–I_{write} loops obtained for various values of I_{write}^{min} (Methods). The Mn_3Sn/Pt device gives a multi-level signal that is tunable according to the magnitude of the write current I_{write}^{min}. This indicates that our antiferromagnetic device

memorizes the amount of electrical current that passes through it—that is, acts as a memristor—and could be useful for neuromorphic computing[43,49].

In recent years, AFMs have attracted considerable attention because they have vanishingly small stray fields perturbing neighbouring cells and much faster spin dynamics than their ferromagnetic counterparts[17,18], leading to successful control of the electrical current in antiferromagnetic sublattices and its detection using anisotropic or spin Hall magnetoresistance[44–47]. However, unlike conventional FM spintronics, these emerging technologies require additional operation schemes to apply currents along directions other than the crystalline axes, hampering their integration to conventional spintronics.

By contrast, our experimental demonstration of room-temperature electrical switching of an antiferromagnetic Weyl metal indicates that the topological AFM may replace the active element of a spintronics device. Because Mn_3Sn has robust topological properties, its polycrystalline form is useful for reading and writing using the same control protocols as those developed for FMs. In addition, the same protocol could be applied not only to Mn_3Sn but also probably to other Mn_3X materials and other AFMs with similar magnetic symmetry[41]. A recent report on the electrical switching behaviour of the antiperovskite Mn_3GaN[50] suggests a similar mechanism.

Finally, our work using the Weyl AFM Mn_3Sn indicates that the SOT switching is a convenient tool to electrically manipulate the distribution of Weyl points and the Berry curvature in momentum space. Thus, our study suggests the prospect of developing a topological antiferromagnetic spintronics field by studying the dynamics and emergent electromagnetism through integrating topological AFMs with spintronics technologies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2211-2.

1. Nielsen, H. B. & Ninomiya, M. The Adler–Bell–Jackiw anomaly and Weyl fermions in a crystal. *Phys. Lett. B* **130**, 389–396 (1983).
2. Wan, X., Turner, A. M., Vishwanath, A. & Savrasov, S. Y. Topological semimetal and Fermi-arc surface states in the electronic structure of pyrochlore iridates. *Phys. Rev. B* **83**, 205101 (2011).
3. Burkov, A. A. & Balents, L. Weyl semimetal in a topological insulator multilayer. *Phys. Rev. Lett*. **107**, 127205 (2011).
4. Armitage, N. P., Mele, E. J. & Vishwanath, A. Weyl and Dirac semimetals in three-dimensional solids. *Rev. Mod. Phys*. **90**, 015001 (2018).
5. Yang, K. Y., Lu, Y. M. & Ran, Y. Quantum Hall effects in a Weyl semimetal: possible application in pyrochlore iridates. *Phys. Rev. B* **84**, 075129 (2011).
6. Son, D. T. & Spivak, B. Z. Chiral anomaly and classical negative magnetoresistance of Weyl metals. *Phys. Rev. B* **88**, 104412 (2013).
7. Zhong, S., Orenstein, J. & Moore, J. E. Optical gyrotropy from axion electrodynamics in momentum space. *Phys. Rev. Lett*. **115**, 117403 (2015).
8. Xiong, J. et al. Evidence for the chiral anomaly in the Dirac semimetal Na_3Bi. *Science* **350**, 413–416 (2015).
9. Nakatsuji, S., Kiyohara, N. & Higo, T. Large anomalous Hall effect in a non-collinear antiferromagnet at room temperature. *Nature* **527**, 212–215 (2015).
10. Ikhlas, M. et al. Large anomalous Nernst effect at room temperature in a chiral antiferromagnet. *Nat. Phys*. **13**, 1085–1090 (2017).
11. Li, X. et al. Anomalous Nernst and Righi–Leduc effects in Mn_3Sn: Berry curvature and entropy flow. *Phys. Rev. Lett*. **119**, 056601 (2017).
12. Kuroda, K. et al. Evidence for magnetic Weyl fermions in a correlated metal. *Nat. Mater*. **16**, 1090–1095 (2017).
13. Nandy, S. et al. Chiral anomaly as the origin of the planar Hall Effect in Weyl semimetals. *Phys. Rev. Lett*. **119**, 176804 (2017).
14. Sakai, A. et al. Giant anomalous Nernst effect and quantum-critical scaling in a ferromagnetic semimetal. *Nat. Phys*. **14**, 1119–1124 (2018).
15. Liu, E. et al. Giant anomalous Hall effect in a ferromagnetic kagome-lattice semimetal. *Nat. Phys*. **14**, 1125–1131 (2018).
16. Kumar, N. et al. Planar Hall effect in the Weyl semimetal GdPtBi. *Phys. Rev. B* **98**, 041103 (2018).
17. Jungwirth, T., Marti, X., Wadley, P. & Wunderlich, J. Antiferromagnetic spintronics. *Nat. Nanotechnol*. **11**, 231–241 (2016).
18. Baltz, V. et al. Antiferromagnetic spintronics. *Rev. Mod. Phys*. **90**, 015005 (2018).
19. Higo, T. et al. Anomalous Hall effect in thin films of the Weyl antiferromagnet Mn_3Sn. *Appl. Phys. Lett*. **113**, 202402 (2018).
20. Yang, H. et al. Topological Weyl semimetals in the chiral antiferromagnetic materials Mn_3Ge and Mn_3Sn. *New J. Phys*. **19**, 015008 (2017).
21. Liu, L., Moriyama, T., Ralph, D. C. & Buhrman, R. A. Spin-torque ferromagnetic resonance induced by the spin Hall effect. *Phys. Rev. Lett*. **106**, 036601 (2011).
22. Pai, C. F. et al. Spin transfer torque devices utilizing the giant spin Hall effect of tungsten. *Appl. Phys. Lett*. **101**, 122404 (2012).
23. Miron, I. M. et al. Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection. *Nature* **476**, 189–193 (2011).
24. Liu, L. et al. Spin-torque switching with the giant spin Hall effect of tantalum. *Science* **336**, 555–558 (2012).
25. Hasan, M. Z. & Kane, C. L. Topological insulators. *Rev. Mod. Phys*. **82**, 3045 (2010).
26. Ando, Y. Topological insulator materials. *J. Phys. Soc. Jpn*. **82**, 102001 (2013).
27. Chien, C. L. & Westgate, C. R. *The Hall Effect and its Applications* (Plenum, 1980).
28. Nagaosa, N., Sinova, J., Onoda, S., MacDonald, A. H. & Ong, N. P. Anomalous Hall effect. *Rev. Mod. Phys*. **82**, 1539–1592 (2010).
29. Xiao, D., Chang, M. C. & Niu, Q. Berry phase effects on electronic properties. *Rev. Mod. Phys*. **82**, 1959–2007 (2010).
30. Machida, Y., Nakatsuji, S., Onoda, S., Tayama, T. & Sakakibara, T. Time-reversal symmetry breaking and spontaneous Hall effect without magnetic dipole order. *Nature* **463**, 210–213 (2010).
31. Chen, H., Niu, Q. & MacDonald, A. H. Anomalous Hall effect arising from noncollinear antiferromagnetism. *Phys. Rev. Lett*. **112**, 017205 (2014).
32. Kiyohara, N., Tomita, T. & Nakatsuji, S. Giant anomalous Hall effect in the chiral antiferromagnet Mn_3Ge. *Phys. Rev. Appl*. **5**, 064009 (2016).
33. Nayak, A. K. et al. Large anomalous Hall effect driven by a nonvanishing Berry curvature in the noncolinear antiferromagnet Mn_3Ge. *Sci. Adv*. **2**, e1501870 (2016).
34. Liu, Z. H. et al. Transition from anomalous Hall effect to topological Hall effect in hexagonal non-collinear magnet Mn_3Ga. *Sci. Rep*. **7**, 515 (2017).
35. Liu, Z. Q. et al. Electrical switching of the topological anomalous Hall effect in a non-collinear antiferromagnet above room temperature. *Nat. Electron*. **1**, 172–177 (2018).
36. Ikeda, T. et al. Anomalous Hall effect in polycrystalline Mn_3Sn thin films. *Appl. Phys. Lett*. **113**, 222405 (2018).
37. Ye, L. et al. Massive Dirac fermions in a ferromagnetic kagome metal. *Nature* **555**, 638–642 (2018).
38. Kimata, M. et al. Magnetic and magnetic inverse spin Hall effects in a non-collinear antiferromagnet. *Nature* **565**, 627–630 (2019); correction **566**, E4 (2019).
39. Liu, J. & Balents, L. Anomalous Hall effect and topological defects in antiferromagnetic Weyl semimetals: Mn_3Sn/Ge. *Phys. Rev. Lett*. **119**, 087202 (2017).
40. Železný, J., Zhang, Y., Felser, C. & Yan, B. Spin-polarized current in noncollinear antiferromagnets. *Phys. Rev. Lett*. **119**, 187204 (2017).
41. Suzuki, M. T., Koretsune, T., Ochi, M. & Arita, R. Cluster multipole theory for anomalous Hall effect in antiferromagnets. *Phys. Rev. B* **95**, 094406 (2017).
42. Higo, T. et al. Large magneto-optical Kerr effect and imaging of magnetic octupole domains in an antiferromagnetic metal. *Nat. Photon*. **12**, 73–78 (2018).
43. Fukami, S., Zhang, C. DuttaGupta, S., Kurenkov, A. & Ohno, H. Magnetization switching by spin-orbit torque in an antiferromagnet-ferromagnet bilayer system. *Nat. Mater*. **15**, 535–541 (2016).
44. Wadley, P. et al. Electrical switching of an antiferromagnet. *Science* **351**, 587–590 (2016).
45. Bodnar, S. Yu. et al. Writing and reading antiferromagnetic Mn_2Au by Néel spin–orbit torques and large anisotropic magnetoresistance. *Nat. Commun*. **9**, 348 (2018).
46. Moriyama, T., Oda, K., Ohkochi, T., Kimata, M. & Ono, T. Spin torque control of antiferromagnetic moments in NiO. *Sci. Rep*. **8**, 14167 (2018).
47. Chen, X. Z. et al. Antidamping-torque-induced switching in biaxial antiferromagnetic insulators. *Phys. Rev. Lett*. **120**, 207204 (2018).
48. Slonczewski, J. C. Current-driven excitation of magnetic multilayers. *J. Magn. Magn. Mater*. **159**, L1–L7 (1996).
49. Olejník, K. et al. Antiferromagnetic CuMnAs multi-level memory cell with microelectronic compatibility. *Nat. Commun*. **8**, 15434 (2017).
50. Hajiri, T., Ishino, S., Matsuura, K. & Asano, H. Electrical current switching of the noncollinear antiferromagnet Mn_3GaN. *Appl. Phys. Lett*. **115**, 052403 (2019).

Methods

Sample and device fabrication

The samples used for the switching measurements, which consist of Ru(2)/Mn_3Sn(40)/Pt(0–14), Ru(2)/Mn_3Sn(40)/W(0–14) or Ru(2)/Mn_3Sn(40)/Cu(0, 5)/AlO_x(5) (parentheses show thicknesses in nanometres) multilayers, are grown on thermally oxidized Si wafers (Fig. 1d). Initially, Ru and Mn_3Sn are deposited at room temperature using a d.c. magnetron sputtering machine with a base pressure of less than 5×10^{-7} Pa (at a rate of ~0.2 nm s^{-1} with a power of 60 W and Ar gas pressure of 0.6 Pa for the Mn_3Sn layer), followed by annealing in situ at 450 °C for 0.5 h. After cooling to room temperature, the Cu, Pt and/or W layers and the AlO_x layer are grown by molecular beam epitaxy under ultrahigh vacuum with a base pressure of less than 2×10^{-8} Pa. We carry out all the fabrication processes in situ, including the sample transfer from the sputter chamber to the molecular beam epitaxy chamber. The composition of the Mn_3Sn layer is $Mn_{3.01(2)}Sn_{0.99(2)}$ (parentheses indicate standard errors), as determined by scanning electron microscopy energy-dispersive X-ray spectrometry (SEM-EDX). Structural analysis is performed with X-ray diffraction using a monochromated Cu Kα source with a wavelength of 1.54 Å. Mn_3Sn is confirmed as a single phase of the $D0_{19}$ structure[51]. For the switching measurements, the Mn_3Sn multilayers are patterned into Hall bar devices with in-plane dimensions of 16 μm × 96 μm using standard photolithography combined with a dry etching process, and Ti/Au contact pads and wires are deposited by electron beam evaporation as shown in Fig. 1d.

Transport measurements

A standard four-probe method with a reading current of I_{read} = 0.2 mA is employed to measure a Hall voltage V_H of the Mn_3Sn Hall bar devices as presented in Fig. 1d. In the SOT-induced switching measurements, we apply a writing pulse current of I_{write} with a duration of 100 ms and then turn it off. Subsequently, we apply I_{read} and start reading V_H after a wait time of 600 ms, as shown in Extended Data Fig. 6a. The estimated heating of the central part of the device is ~50 K, and it remains lower than the Néel point even when the write current I_{write} is on. Moreover, V_H is not affected by the heating caused by the I_{write} injection, but represents the Hall voltage value at each monitored temperature, because the wait time of 600 ms between the I_{write} injection and the V_H measurement (Extended Data Fig. 6a) is long enough, as confirmed by the wait time dependence of V_H (Extended Data Fig. 7d). Except for the low-temperature results presented in Fig. 2e, all measurements are performed at room temperature (~293 K). To estimate the current density in the NM, the resistivity of Ru/Mn_3Sn, Pt and W are calculated on the basis of the Si/SiO_2/Ru(2)/Mn_3Sn(40)/AlO_x(5) and Si/SiO_2/Ru(2)/Mn_3Sn(40)/Pt or W(d_{NM})/AlO_x(5) Hall bar devices using the 'conductors in parallel' model. A two-probe method with contact pads for the current is used for this resistivity measurement. The thickness dependence of the resistivity in the NM layer is shown in Extended Data Fig. 6b. The resistivity of Mn_3Sn(40) thin films deposited with the same method has been reported[19] to be ~300 μΩ cm, consistent with measurement results for a bulk crystal[9]. The electrical measurement presented in Fig. 2e is performed under 6×10^{-5} Pa in a vacuum chamber cooled by a helium compressor. The sample substrate is fixed on a Cu sample stage and the temperature is measured by a thermometer inside the sample stage. For the V_H–I_{write} loop measurements presented in Fig. 3b, we first initialize the magnetic domains by applying a positive pulse current I_{ini} larger than the switching current I_c. Then, we scan the current I_{write} from I_{ini} to I_{write}^{min} under H_x = +0.1 T. The Mn_3Sn/Pt device exhibits a multi-stable signal according to the magnitude of the write current I_{write}^{min}—that is, acts as a memristor—and thus would be of great importance for neuromorphic computing[43,49,52,53].

For the longitudinal magnetoconductivity (MC) and planar Hall effect (PHE) measurements, a 40-nm-thick Mn_3Sn layer deposited directly on the Si/SiO_2 substrates is fabricated using the same synthesis conditions as those used for the multi-layer devices used for the switching measurements. As shown in Extended Data Figs. 1e, 2c, bar-shaped samples with voltage terminals made of Ag paste are measured. To obtain the magnetic field angular dependence, the horizontal rotator option of a commercial physical property measurement system (PPMS) is used. Similarly to the MC and PHE measurements, bar-shaped samples (9 mm length, 2 mm width) are used in the PPMS to measure the transport properties—including the resistivity and Hall, Seebeck and Nernst effects—presented in Extended Data Fig. 4. The electrical and thermal currents are applied along the length direction, and the magnetic field is applied along the perpendicular direction to the film plane. The transverse thermoelectric conductivity α_N is estimated from the electrical conductivity σ, the Hall conductivity σ_H, the Seebeck coefficient S_{SE} and the Nernst coefficient S_{ANE} using the equation[29] $\alpha_N = \sigma S_{ANE} + \sigma_H S_{SE}$.

Experimental evidence of Weyl metal state in polycrystalline Mn_3Sn thin film

To demonstrate the switching of the topological antiferromagnetic state, we first need to show that the polycrystalline sample does support the same Weyl semimetal state as has been found in Mn_3Sn single crystals[10,12,20].

Evidence for the presence of Weyl fermions can be obtained using various probes, such as: (1) angle-resolved photoemission spectroscopy, to reveal the band structure; (2) detection of the chiral anomaly through the MC and PHE; (3) (when the Weyl fermions are due to TRS breaking) observation of a large ANE beyond the empirical scaling law with magnetization, which is known to provide further strong evidence for Weyl fermions[10,11,14,54].

In the case of the bulk single crystal of Mn_3Sn, all of the three probes have been used to establish the magnetic Weyl semimetal state[10,12,20]. To further obtain experimental evidence for the switching of the topological AFM and the manipulation of the distribution of the Weyl points in our polycrystalline thin films, we employ the following approach using the transport probes (2) and (3), which are both useful for polycrystalline samples. First, the observation of the chiral anomaly should provide strong evidence for Weyl fermions (see, for example, refs. [1,4,6,8,12,13,16,55–58]). Second, a large ANE—much larger than what would be expected according to the empirical scaling law with magnetization—should provide further strong evidence for the presence of Weyl points near the Fermi energy, which produces a large Berry curvature[10,11,14,54,59]. Third, because the ANE probes not only the size but also the sign of the Berry curvature at the Fermi energy[29], the motion of the Weyl points that produce the Berry curvature should be detected as a sign change in the ANE. Thus, the combination of the chiral anomaly and the sign change in a large ANE should provide clear evidence for the motion of the Weyl points near the Fermi energy in momentum space.

Chiral anomaly

Positive longitudinal magnetoconductivity. As mentioned, the chiral anomaly has been established as evidence for Weyl fermions (see, for example, refs. [1,4,6,8,12,13,16,55–58]). The chiral anomaly refers to positive longitudinal MC due to the violation of a separate number conservation law of three-dimensional left- and right-handed Weyl fermions. The anomaly is known to arise when the magnetic field **H** is aligned parallel to the electric field **E**. Because this effect depends on the scalar product $|\mathbf{E}\cdot\mathbf{H}|$, the qualitative features of the MC should mostly depend on the angle between **E** and **H** and should appear in a polycrystalline sample. Thus, if Weyl fermions are present in the polycrystalline form of Mn_3Sn, the chiral anomaly should lead to an anisotropic MC—namely, positive longitudinal MC—when the magnetic field is applied parallel to the electric field ($\mathbf{H} \parallel \mathbf{E}$), whereas negative transverse MC should appear when the magnetic field is perpendicular to the electrical current ($\mathbf{H} \perp \mathbf{E}$).

To examine whether the chiral anomaly is present in our polycrystalline films, we carry out magneto-transport measurements. We find that the anisotropic field dependence of the MC is fully consistent with the

chiral anomaly, as shown in Extended Data Fig. 1a. Namely, the longitudinal MC for a magnetic field parallel to the current ($\mathbf{H} \parallel \mathbf{E}$, $\Theta_{ch} = 0°$) increases with $|\mathbf{H}|$. On the other hand, a magnetic field perpendicular to the electric field ($\mathbf{H} \perp \mathbf{E}$, $\Theta_{ch} = 90°$) generates a much smaller but finite increase with $|\mathbf{H}|$, which is attributable to the suppression of spin fluctuations. Generally, in magnetic conductors the application of the magnetic field may lead to the suppression of thermally induced spin fluctuations, and thus positive MC, irrespective of the angle Θ_{ch} between $\mathbf{H}$ and $\mathbf{E}$. Although this effect may complicate the detection of the chiral anomaly, this type of positive longitudinal MC should be reduced upon cooling. In sharp contrast, we find that the positive longitudinal MC for $\mathbf{H} \parallel \mathbf{E}$ increases by nearly 50% by decreasing the temperature from 300 K to 250 K (Extended Data Fig. 1a–d). This indicates that the longitudinal MC does not originate from the suppression of spin fluctuations, but from the chiral anomaly itself. In addition, the transverse MC for $\mathbf{H} \perp \mathbf{E}$ becomes negative under fields higher than 5 T (Extended Data Fig. 1b), indicating that the effect of spin fluctuations can be suppressed under high fields at 250 K. Thus, our experiment indicates that the main origin of anisotropic MC is the chiral anomaly both at 250 K and at 300 K.

Another conventional mechanism that may yield positive longitudinal MC is the current-jetting effect. In high-mobility semimetals, the orbital effect leads to a large negative transverse MC compared with the longitudinal MC[4,60–62]. This anisotropic MC may orient the current to the direction parallel to the magnetic field and suppress the transverse component of the current flow, resulting in a positive longitudinal MC that strongly depends on the sample geometry. Recent studies have revealed that the positive longitudinal MC observed in several Dirac and Weyl semimetals comes from the current-jetting effect[4,62]. This effect is unlikely to occur in Mn_3Sn because both the mobility and the anisotropy in the MC are much lower than those reported for the above semimetals. Nevertheless, we examine the possibility of current inhomogeneity by preparing three pairs of voltage terminals, which we place on the centre line and on both sides of a Mn_3Sn polycrystalline film (Extended Data Fig. 1e). All three voltage probes detect the same field dependence of the longitudinal MC and transverse MC, confirming that the current distribution is homogeneous inside the film and ruling out the current-jetting effect as the origin of the positive longitudinal MC (Extended Data Fig. 1a).

Planar Hall effect. Recent theory has shown that the chiral anomaly may lead to not only anisotropic MC (positive longitudinal MC) but also PHE[13,16,57,58]. The conductivity σ and planar Hall conductivity σ_H^{PHE} are formulated as

$$\sigma = \sigma_{\perp} + \Delta\sigma_{chiral}\cos^2\Phi_{ch} \quad (2)$$

$$\sigma_H^{PHE} = \Delta\sigma_{chiral}\sin\Phi_{ch}\cos\Phi_{ch} \quad (3)$$

Here $\Delta\sigma_{chiral}$ ($\Delta\sigma_{chiral} = \sigma_{\parallel} - \sigma_{\perp}$) is the positive MC induced by the chiral anomaly, where $\sigma_{\parallel}$ and $\sigma_{\perp}$ is the MC when the electrical current is parallel ($\Phi_{ch} = 0°$) and perpendicular ($\Phi_{ch} = 90°$) to the magnetic field, respectively, and Φ_{ch} is the in-plane angle between the magnetic field and the electrical current, as shown in Extended Data Fig. 2c.

Figure 1c and Extended Data Figs. 2a, 2b present the Φ_{ch} dependence of the MC $\Delta\sigma = \sigma - \sigma_{\perp}$ and the planar Hall conductivity σ_H^{PHE}. As noted above, the electrical current direction and the sample are rotated in the plane with an angle of Φ_{ch} from the magnetic field direction (Extended Data Fig. 2c). The anisotropic MC $\Delta\sigma$ exhibits a sinusoidal dependence on Φ_{ch}. Our results obtained at 300 K (Fig. 1c) and 250 K (Extended Data Figs. 2a, 2b) for both anisotropic MC and PHE are fully consistent with the theory and well fitted by equations (2) and (3). In further agreement with equations (2) and (3), the same $|\Delta\sigma_{chiral}|$ values (~0.4 Ω^{-1} cm^{-1} at 300 K, ~1 Ω^{-1} cm^{-1} at 250 K) are obtained for both $\Delta\sigma(\Phi_{ch})$ and $\sigma_H^{PHE}(\Phi_{ch})$.

To summarize, our MC and PHE measurements provide strong evidence for the chiral anomaly and thus for the presence of Weyl fermions in polycrystalline Mn_3Sn thin films.

Large ANE

Further evidence for the presence of Weyl fermions can be obtained from ANE measurements. The ANE is usually found in FMs and is empirically known to scale with magnetization[10]. In magnetic Weyl semimetals, however, Weyl points near the Fermi energy should generate a large net Berry curvature, leading to a large ANE beyond this empirical scaling law[10,11,14,54]. A prominent example has been found in the case of the Mn_3Sn single crystals, as shown in Extended Data Fig. 3. Namely, despite its vanishingly small magnetization (~0.003μ_B per Mn atom), Mn_3Sn exhibits a large ANE ($S_{ANE} \approx 0.35\ \mu V\ K^{-1}$), equivalent to those obtained for FMs with a large magnetization of ~1μ_B according to the empirical scaling law (Extended Data Fig. 3)[10].

To confirm the large ANE driven by Weyl fermions, we measured the ANE of our polycrystalline Mn_3Sn thin film. The field dependence of the anomalous Nernst coefficient S_{ANE} and the Hall resistivity ρ_H is presented in Extended Data Fig. 4a. The polycrystalline Mn_3Sn film exhibits a large spontaneous ANE with $S_{ANE} \approx 0.25\ \mu V\ K^{-1}$ at 300 K, similar to single-crystal values. The spontaneous magnetization of the film is found to be only ~0.006μ_B per Mn atom[19], again similar to the values observed for single crystals. Thus, the S_{ANE} value obtained for the film is more than two orders of magnitude larger than that expected from the empirical scaling law with magnetization (Extended Data Fig. 3).

Generally, the ANE has two components, $S_{ANE} = \alpha_N/\sigma - S_{SE}(\sigma_H/\sigma)$, where α_N is the transverse thermoelectric coefficient, σ is the electrical conductivity and S_{SE} is the Seebeck coefficient. If the ANE is enhanced by Weyl points near the Fermi energy E_F, the first term should be dominant. This is because the Berry curvature $\Omega_{n,z}(\mathbf{k})$ near E_F, which is enhanced by the presence of Weyl points near E_F, determines the thermoelectric coefficient α_N, as described by

$$\alpha_N \approx \sum_{n,\mathbf{k}} \Omega_{n,z}(\mathbf{k})\delta(E_F - \varepsilon_{n,\mathbf{k}}) \quad (4)$$

whereas the anomalous Hall conductivity represents the sum of the Berry curvatures of all the occupied bands[29]. In equation (4), n, $\mathbf{k}$, and $\varepsilon_{n,\mathbf{k}}$ is the band index, the wave vector and the band energy, respectively.

To clarify the origin of the large ANE, we performed an analysis and found that the transverse thermoelectric coefficient $\alpha_N = \sigma S_{ANE} + \sigma_H S_{SE} \approx 0.1\ A\ K^{-1}\ m^{-1}$ at 300 K (Extended Data Fig. 4b) is ~60% of the value obtained for bulk single crystals[10] (see Methods section 'Transport measurements' and the caption of Extended Data Fig. 4 for the estimation of α_N). We note that the first term ($\sigma S_{ANE} \approx 0.09\ A\ K^{-1}\ m^{-1}$) is much larger than the second term ($\sigma_H S_{SE} \approx 0.01\ A\ K^{-1}\ m^{-1}$), namely, $\alpha_N \approx \sigma S_{ANE}$, which indicates that the large α_N due to the large Berry curvature near E_F dominates S_{ANE}. The large spontaneous S_{ANE} and α_N provide further strong evidence for the presence of Weyl fermions in the Mn_3Sn polycrystalline film.

Sign change in ANE as evidence of switching of the topological AFM

In a magnetic Weyl semimetal, the breaking of the TRS generates a pair of Weyl points in momentum space, and the symmetry constraint requires the nodal direction connecting a pair of Weyl points to be aligned along the TRS-breaking field or order parameter. Thus, reversing the magnetic field direction or order parameter rotates the pairs of Weyl points by 180° in momentum space. In Mn_3Sn, the reversal of the cluster octupole polarization flips the pairs of Weyl points and the associated Berry curvatures, which can be viewed as an axial vector[12], as illustrated in Extended Data Fig. 5.

Equation (4) indicates that such rotation of the Berry curvature should lead to a sign change in α_N. As discussed, α_N is highly enhanced

for the polycrystalline Mn_3Sn thin film owing to the presence of Weyl points near the Fermi energy, as in the case of single crystals. In such a magnetic Weyl semimetal, therefore, a sign change in α_N observed during the field sweep should come from the 180° rotation of the pairs of Weyl points near E_F. Moreover, as shown in Extended Data Fig. 4b, α_N and σ_H show a sign change by reversing the magnetic field, and their field dependences overlap in the case of the thin film, further indicating that the sign change in σ_H should also come from the flipping of the pairs of Weyl points. Thus, we conclude that the jump in the AHE signal (V_H) induced by electrical switching probes the rotation of the pairs of Weyl points and the associated Berry curvature in the Mn_3Sn thin films. In particular, for the model shown in Fig. 4, the switching should take place between the two different configurations of Weyl points shown in Extended Data Fig. 5a and Extended Data Fig. 5b.

In summary, all our measurements and analyses demonstrate that the Weyl physics found in bulk Mn_3Sn is also applicable to polycrystalline thin films. Thus, the electrical switching of the AHE signal in the Mn_3Sn thin film indicates electrical manipulation of the topological antiferromagnetic state and of the distribution of Weyl points in momentum space.

LLG equation based on one-layer kagome lattice Hamiltonian

To understand the deterministic switching mechanism based on SOT, we study the dynamics of the sublattice moments $\mathbf{m}_{ia}$ that obeys the LLG equation[21,48] using the following Hamiltonian defined on the one-layer kagome lattice (blue layer on the y–z plane in Fig. 4a)

$$\mathcal{H} = \mathcal{J}\sum_{\langle ia,jb\rangle}\mathbf{m}_{ia}\cdot\mathbf{m}_{jb} + \mathcal{D}\sum_{\langle i,j\rangle}\mathbf{x}\cdot(\mathbf{m}_{i1}\times\mathbf{m}_{j2}+\mathbf{m}_{i2}\times\mathbf{m}_{j3}+\mathbf{m}_{i3}\times\mathbf{m}_{j1}) - \mathcal{K}\sum_{ia}(\mathbf{k}_a\cdot\mathbf{m}_{ia})^2 \quad (5)$$

where the suffixes i and j denote the unit cell, a and b = (1, 2, 3) are the sublattices, and $\mathbf{x}$ = (1, 0, 0) is the unit vector along the x axis (Fig. 4a). We note that this model includes only three Mn atoms in a unit cell, and we neglect inter-layer couplings because they do not give qualitatively new effects[39]. $\mathbf{m}_{ia}$ denotes a unit magnetic moment. The nearest-neighbour exchange interaction $\mathcal{J}$, the Dzyaloshinskii–Moriya (DM) interaction $\mathcal{D}$ and the in-plane magnetic anisotropy $\mathcal{K}$ are assumed to be positive, and they stabilize the ITS texture observed in Mn_3Sn[51,63]. $\mathbf{k}_a = (0, \cos\varphi_a, \sin\varphi_a)$ with $(\varphi_1, \varphi_2, \varphi_3) = (\pi, 9\pi, 5\pi)/6$ lifts the in-plane U(1) degeneracy and fixes the sixfold symmetry, where φ_a is the angle of the sublattice moments (a = 1, 2, 3) with respect to the y axis in the kagome plane (y–z plane in Fig. 4a). Here, $\mathcal{J}$ and $\mathcal{D}$ are identical to $\mathcal{J}S^2$ and $\mathcal{D}S^2$ employed in ref. [64] (and in S.M. et al., manuscript in preparation) and a different definition is adapted for the in-plane anisotropy $\mathcal{K}$, leading to a sixfold magnetocrystalline anisotropy and spontaneous magnetization. The bias field is included in a Zeeman coupling term, $H_{Zeeman} = -\mu_0 M_S \sum_{ia}\mathbf{m}_{ia}\cdot\mathbf{H}$, where $M_S = 3\mu_B$ is the saturation magnetic moment of the Mn atom, and the SOT causes the spin-transfer-induced in-plane torque $\mathbf{T}_{ia} = \frac{\hbar|\gamma|J_0\theta_{SH}}{2em_{SD}}\mathbf{m}_{ia}\times(\mathbf{y}\times\mathbf{m}_{ia})$. On this basis, the LLG equation becomes equation (1)[21,48].

For the simulation of the magnetization switching process in configuration (a), we set parameters $\mathcal{J}$ = 23 meV, $\mathcal{D}$ = 1.6 meV, $\mathcal{K}$ = 0.17 meV and α = 0.003. These are consistent with the experimentally obtained physical parameters in Mn_3Sn, where the sixfold magnetocrystalline anisotropy energy and the spontaneous magnetization are 310 J m^{-1} and 0.01μ_B per f.u., respectively (S.M. et al., manuscript in preparation). The other parameters are set as: θ_{SH} = 0.1, d = 40 nm, $\mu_0|\mathbf{H}|$ = 0.1 T and $J_0 = 6\times10^{14}$ A m^{-2}.

Crystal grain configurations of polycrystalline Mn_3Sn Hall bar devices

When we consider the possible SOT mechanism for the non-collinear antiferromagnetic spin texture in Mn_3Sn using the accumulated spin at the Mn_3Sn/NM interface, the polycrystalline character of our Mn_3Sn layer suggests that there are three crystal grain configurations representing the Hall bar devices: (a) the kagome layer is perpendicular to the current I and parallel to the electrically injected carrier spin polarization $\mathbf{p}$ (Fig. 4a); (b) the kagome layer is parallel to I and perpendicular to $\mathbf{p}$ (Extended Data Fig. 8a); and (c) the kagome layeris parallel to I and parallel to $\mathbf{p}$ (Extended Data Fig. 8b).

As discussed in the main text, in configuration (a) the combination of the spin current and the bias field may cause the deterministic SOT switching (Fig. 4). By contrast, for grains with configuration (b), the spin current injection causes a steady state with continuous rotation of a slightly canted ITS structure, destabilizing the exchange interaction and DM interaction energy. Extended Data Fig. 8c and its inset show an example of a numerical simulation with the spin injection starting at time t = 21.5 ns. We find that the in-plane angle φ (as defined in Fig. 4c) oscillates at t > 21.5 ns with a frequency of ~3.5 THz as a result of the continuous rotation. This indicates that a random configuration of the octupole moments arises in each grain, which depends on the time when the current is turned off[64,65], and thus the final state is not uniquely determined by the bias field. In configuration (c) (Extended Data Fig. 8b), such a grain does not have any out-of-plane (z) component of the octupole moment and thus does not exhibit any AHE.

We note that configuration (a) has two characteristic arrangements of the kagome layer orientations: (a-1) b axis [01$\bar{1}$0] parallel to y (used in the theoretical calculation; see Fig. 4a); (a-2) a axis [2$\bar{1}\bar{1}$0] parallel to y, as shown in Extended Data Fig. 9a, b. In the field-induced switching, both arrangements should contribute to the AHE. In the electrical switching, however, only arrangement (a-1) (Extended Data Fig. 9a) may contribute to the AHE because the octupole polarization (orange arrows) in arrangement (a-2) (Extended Data Fig. 9b) is pinned along its easy axis (a axis, [2$\bar{1}\bar{1}$0]). Thus, some grains with configuration (a) may contribute to the electrical switching.

Deterministic switching in numerical simulations

Here we discuss the simulation results regarding the deterministic switching. As mentioned in the main text, irrespective of the sign of I_{write}, we observe the same steady states in the switching process. This behaviour is counterintuitive, because it contrasts the analogy of conventional FMs, where the magnetic moments are aligned along the direction of $\mathbf{T}_{ia}$, which depends on the sign of I_{write}. Here we show that the damping torque combined with the SOT, $\mathbf{T}'_{ia} = \alpha\mathbf{m}_{ia}\times\mathbf{T}_{ia}$, is key to obtaining this behaviour.

Extended Data Figure 10a shows the evolution of the in-plane motion of the octupole polarization without the bias field, where the onset of the pulse current injection is set at t_1. Whereas both positive and negative I_{write} values lead to the same steady states after long enough time, we note that in the first short period immediately after the onset time t_1, the octupole polarization actually rotates in the directions determined by $\mathbf{T}_{ia}$, depending on the sign of I_{write} (inset of Extended Data Fig. 10a). In particular for the I_{write} < 0 case, we find that the in-plane angle φ returns to its initial value at t_2 (>t_1). We stress that the in-plane components of each sublattice moment at t_1 and t_2 are almost identical in this case. However, the out-of-plane components of the sublattice moments that are induced by $\mathbf{T}'_{ia} \approx \alpha\mathbf{m}_{ia}\times\mathbf{T}_{ia}$ are different, as depicted in Extended Data Fig. 10b. This result indicates that a type of non-adiabatic torque emerges from $\mathbf{T}'_{ia}$ that acts on the spins independently of the sign of I_{write}. Namely, the amplitude of this torque gradually increases with increasing the staggered out-of-plane components induced by $\mathbf{T}'_{ia}$ and finally overcomes the I_{write}-sign-dependent $\mathbf{T}_{ia}$ for $t > t_2$. We note that a usual low-energy effective model, as discussed in refs. [39,64], cannot capture the above feature because it includes only the in-plane motion of the order parameter.

Although the microscopic switching mechanism is different, we can use for Mn_3Sn the same conventional spintronics setup as that used in ferromagnetic devices because the switching conditions are the same. This is achieved with the help of the Gilbert damping torque combined with the SOT and the staggered out-of-plane components of the sublattice moments.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

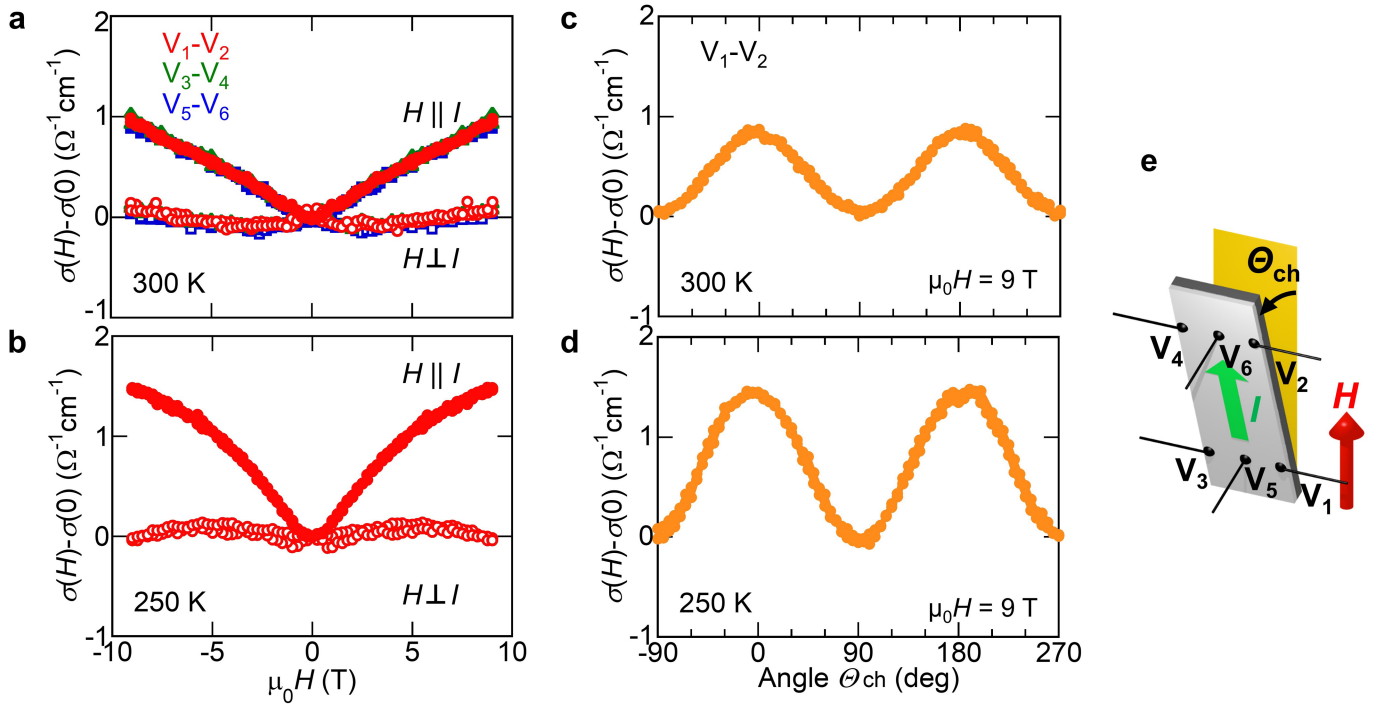
51. Tomiyoshi, S. & Yamaguchi, Y. Magnetic structure and weak ferromagnetism of Mn_3Sn studied by polarized neutron diffraction. *J. Phys. Soc. Jpn*. **51**, 2478–2486 (1982).
52. Lequeux, S. et al. A magnetic synapse: multilevel spin-torque memristor with perpendicular anisotropy. *Sci. Rep*. **6**, 31510 (2016).
53. Kurenkov, A. et al. Artificial neuron and synapse realized in an antiferromagnet/ferromagnet heterostructure using dynamics of spin–orbit torque switching. *Adv. Mater*. **31**, 1900636 (2019).
54. Guin, S. N. et al. Zero-field Nernst effect in a ferromagnetic kagome-lattice Weyl-semimetal $Co_3Sn_2S_2$. *Adv. Mater*. **31**, 1806622 (2019).
55. Zyuzin, A. A. & Burkov, A. A. Topological response in Weyl semimetals and the chiral anomaly. *Phys. Rev. B* **86**, 115133 (2012).
56. Hirschberger, M. et al. The chiral anomaly and thermopower of Weyl fermions in the half-Heusler GdPtBi. *Nat. Mater*. **15**, 1161–1165 (2016).
57. Burkov, A. A. Giant planar Hall effect in topological metals. *Phys. Rev. B* **96**, 041110 (2017).
58. Li, H. et al. Giant anisotropic magnetoresistance and planar Hall effect in the Dirac semimetal Cd_3As_2. *Phys. Rev. B* **97**, 201110 (2018).
59. Sharma, G., Goswami, P. & Tewari, S. Nernst and magnetothermal conductivity in a lattice model of Weyl fermions. *Phys. Rev. B* **93**, 035116 (2016).
60. Pippard, A. B. *Magnetoresistance in Metals* (Cambridge Univ. Press, 1989).
61. Hu, J. et al. Current jets, disorder, and linear magnetoresistance in the silver chalcogenides. *Phys. Rev. Lett*. **95**, 186603 (2005).
62. dos Reis, R. D. et al. On the search for the chiral anomaly in Weyl semimetals: the negative longitudinal magnetoresistance. *New J. Phys*. **18**, 085006 (2016).
63. Nagamiya, T., Tomiyoshi, S. & Yamaguchi, Y. Triangular spin configuration and weak ferromagnetism of Mn_3Sn and Mn_3Ge. *Solid State Commun*. **42**, 385–388 (1982).
64. Nomoto, T. & Arita, R. Cluster multipole dynamics in non-collinear antiferromagnets. *Phys. Rev. Res*. **2**, 012045 (2020).
65. Fujita, H. Field-free, spin-current control of magnetization in non-collinear chiral antiferromagnets. *Phys. Status Solidi Rapid Res. Lett*. **11**, 1600360 (2017).

Acknowledgements We thank D. Qu, T. Tomita, Y. Hibino, T. Nozaki and S. Yuasa for discussions, and D. Nishio-Hamane for SEM-EDX measurements. This work is partially supported by CREST (JPMJCR18T3), Japan Science and Technology Agency (JST), through Grants-in-Aid for Scientific Research on Innovative Areas (15H05882 and 15H05883) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, by Grants-in-Aid for Scientific Research (16H06345, 18H03880, 19H00650) and by the New Energy and Industrial Technology Development Organization.

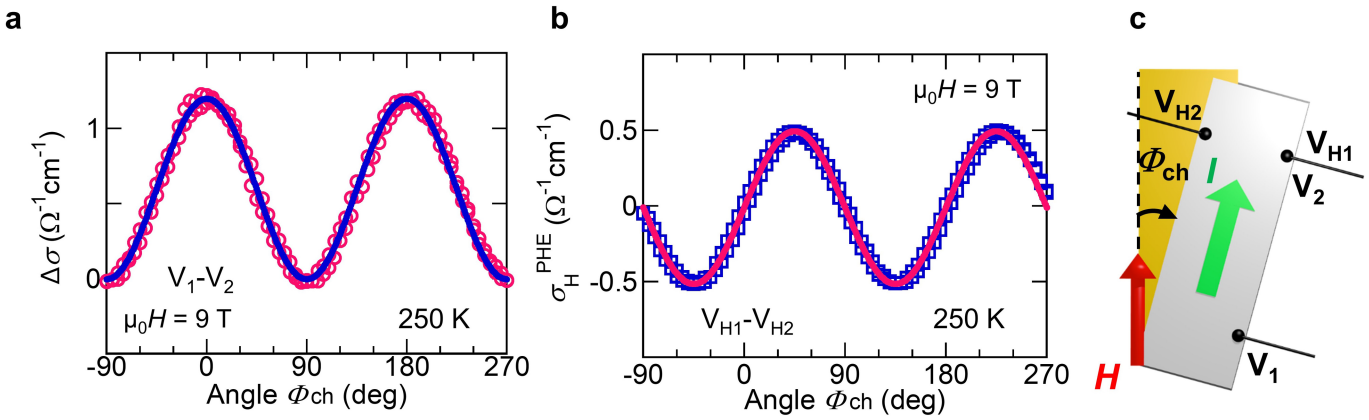
Author contributions S.N. conceived the project. S.N., K.K., S.M. and Y.O. planned the experiments. T.H., S.M., A.K., T. Nakano and K.Y. prepared and characterized the Mn_3Sn multilayered films. K.K. fabricated the Hall bar devices. H.T., T.H., K.K. and S.M. performed the electrical switching measurements. T.H. performed the magneto-transport measurements and A.S. performed the thermoelectric measurements. T. Nomoto and R.A. performed numerical calculations and provided a theoretical explanation. T.H., T. Nomoto, S.M. and S.N. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

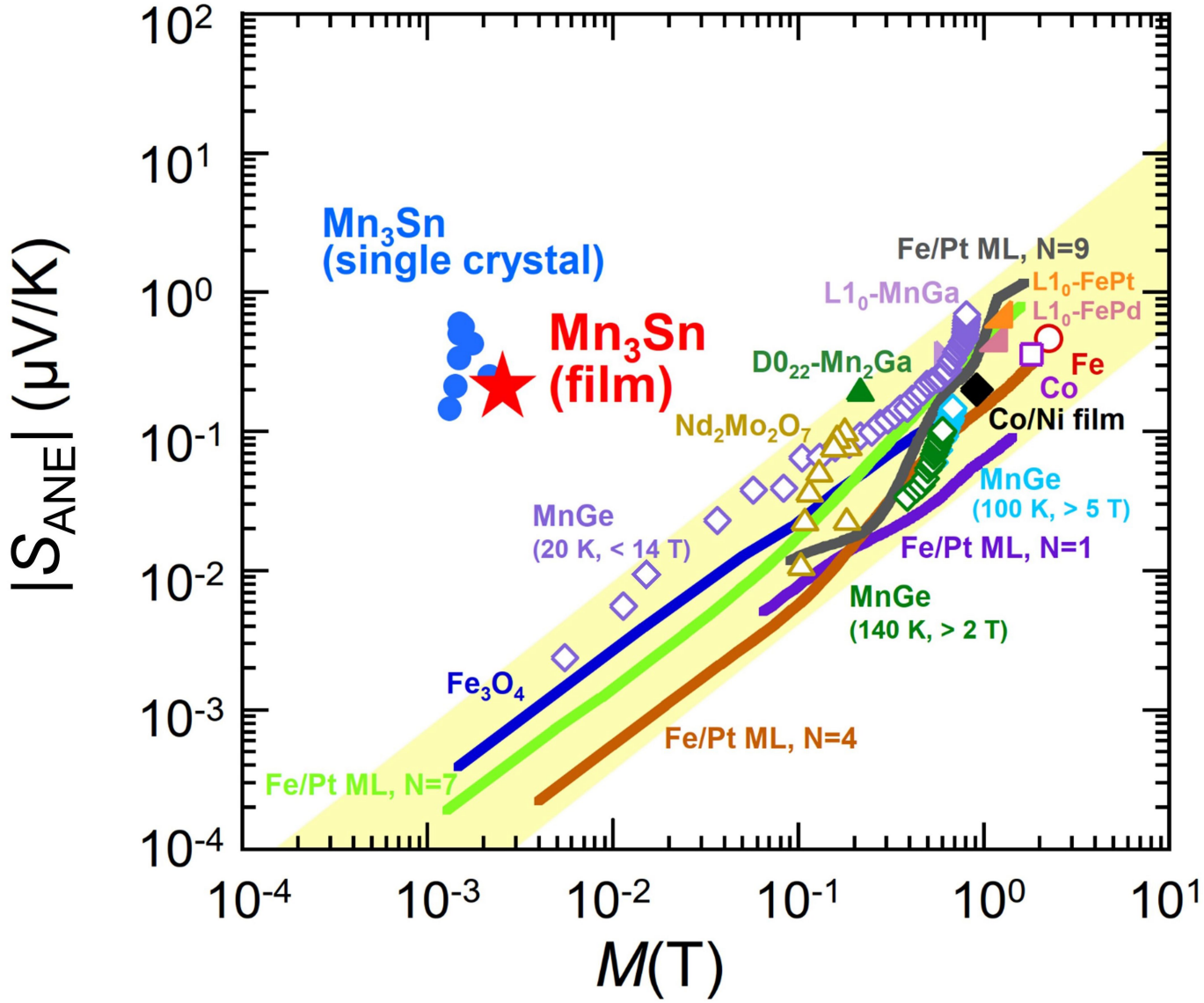
Additional information
Correspondence and requests for materials should be addressed to S.N.
Peer review information *Nature* thanks Kyung-Jin Lee and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



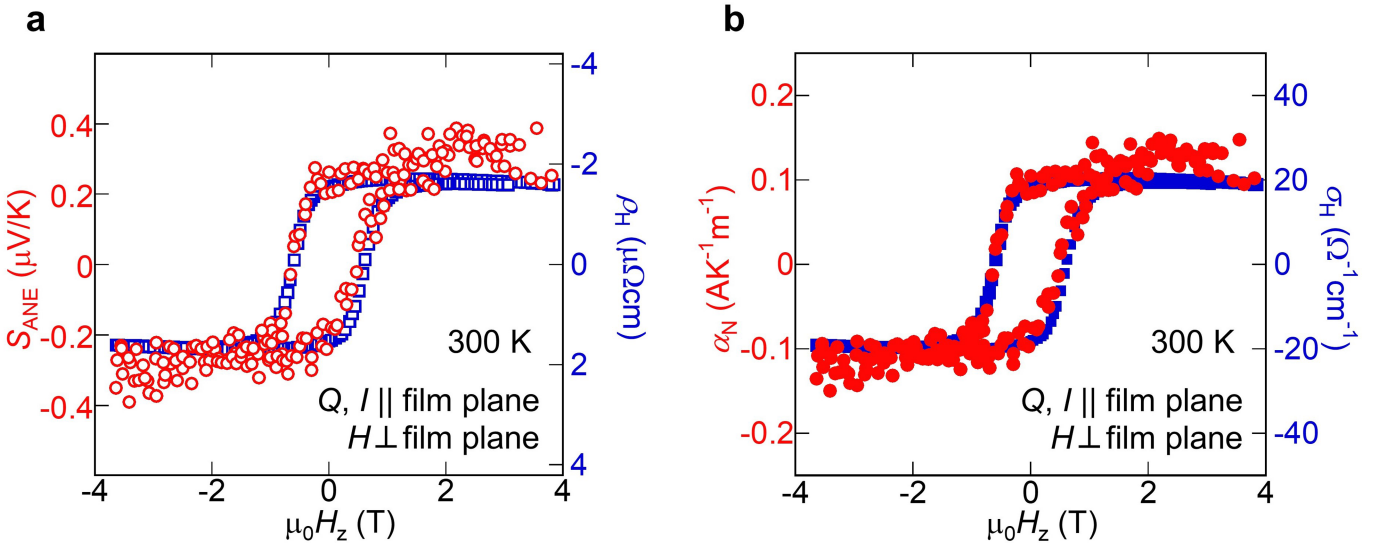
Extended Data Fig. 1 | Longitudinal magnetoconductivity measurements for the Mn_3Sn thin film. a, b, Magnetic field dependence of the magnetoconductivity $\sigma(H) - \sigma(0)$ of the 40-nm-thick Mn_3Sn film under a magnetic field H parallel (top curves) and perpendicular (bottom curves) to the current I at 300 K (**a**) and 250 K (**b**). Here $\sigma(0)$ is the magnetoconductivity at 0 T. **c, d**, Angular (Θ_{ch}) dependence of the magnetoconductivity $\sigma(H) - \sigma(0)$ of the 40-nm-thick Mn_3Sn film at 300 K (**c**) and 250 K (**d**). **e**, Schematic of the experimental setup used for the magneto-transport measurements. To examine the current homogeneity, we employ three pairs of voltage terminals—(V_1, V_2) and (V_3, V_4) on the two sides and (V_5, V_6) on the centre line of the film—for the measurement shown in **a**. For the measurements shown in **b**–**d**, the pair (V_1, V_2) is used.



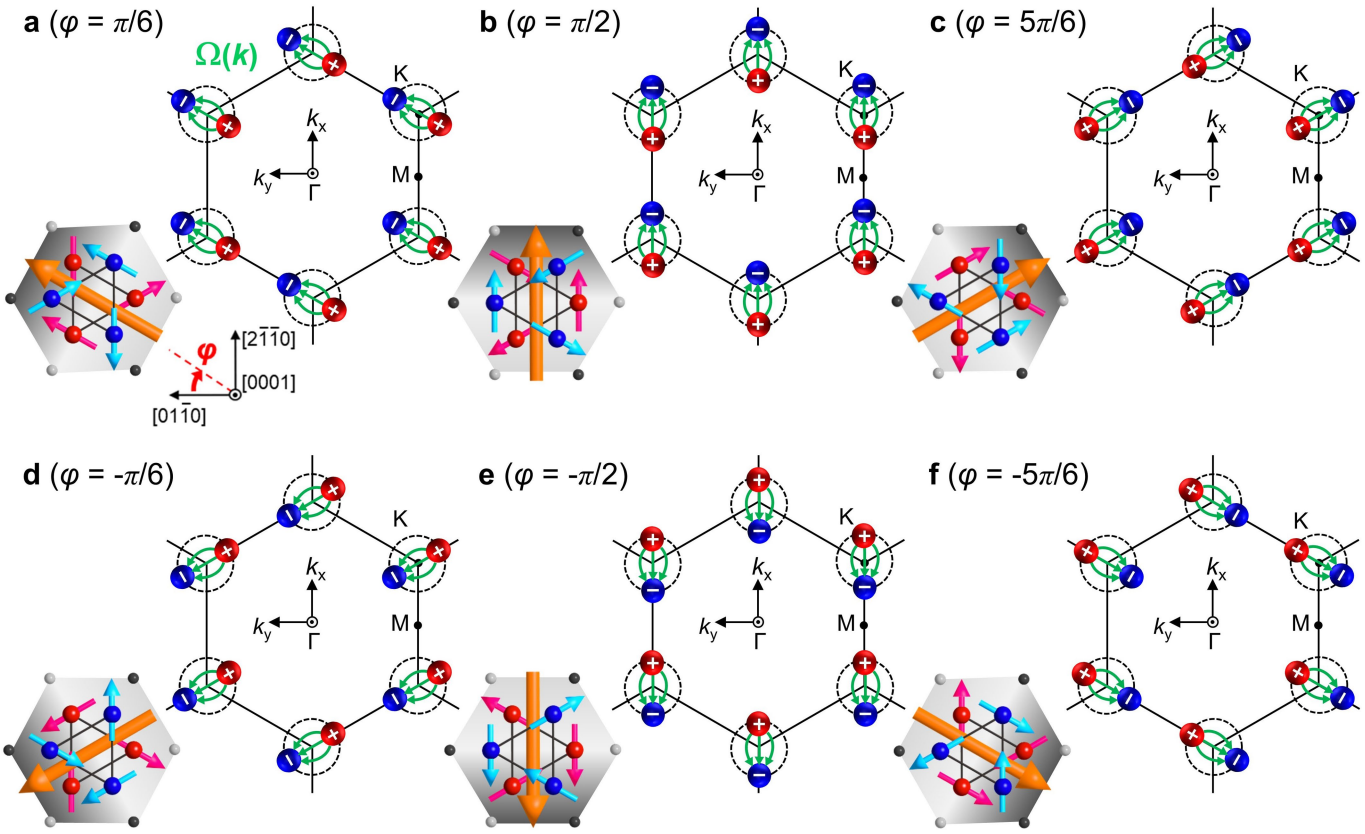
Extended Data Fig. 2 | Longitudinal MC and planar Hall conductivity measurements for the Mn_3Sn thin film. a, Angular (Φ_{ch}) dependence of the longitudinal MC $\Delta\sigma = \sigma - \sigma_{\perp}$ of the 40-nm-thick Mn_3Sn film at 250 K. Φ_{ch} is the in-plane angle between the magnetic field H and the electrical current I. The blue solid line shows the fitting results using equation (2). **b**, Angular (Φ_{ch}) dependence of the planar Hall conductivity σ_H^{PHE} of the 40-nm-thick Mn_3Sn film at 250 K. The pink solid line shows the fitting results using equation (3). **c**, Schematic of the measurement setup used for the longitudinal MC and PHE with the pairs of terminals (V_1, V_2) and (V_{H1}, V_{H2}).



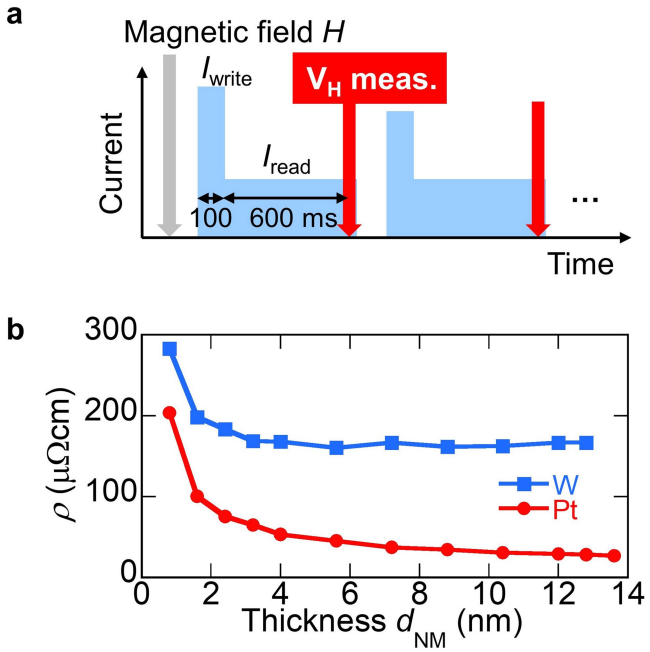
Extended Data Fig. 3 | Large ANE in the Mn_3Sn thin film. Double-logarithmic plot of the anomalous Nernst coefficient $|S_{ANE}|$ versus magnetization M for various FMs, for Mn_3Sn single crystals at various temperatures (blue solid circles) and for the polycrystalline Mn_3Sn thin film at 300 K (red star). The yellow shaded region highlights the empirical scaling law with M. ML, multilayer; N, number of the stacking. Data from ref. [10], Springer Nature.



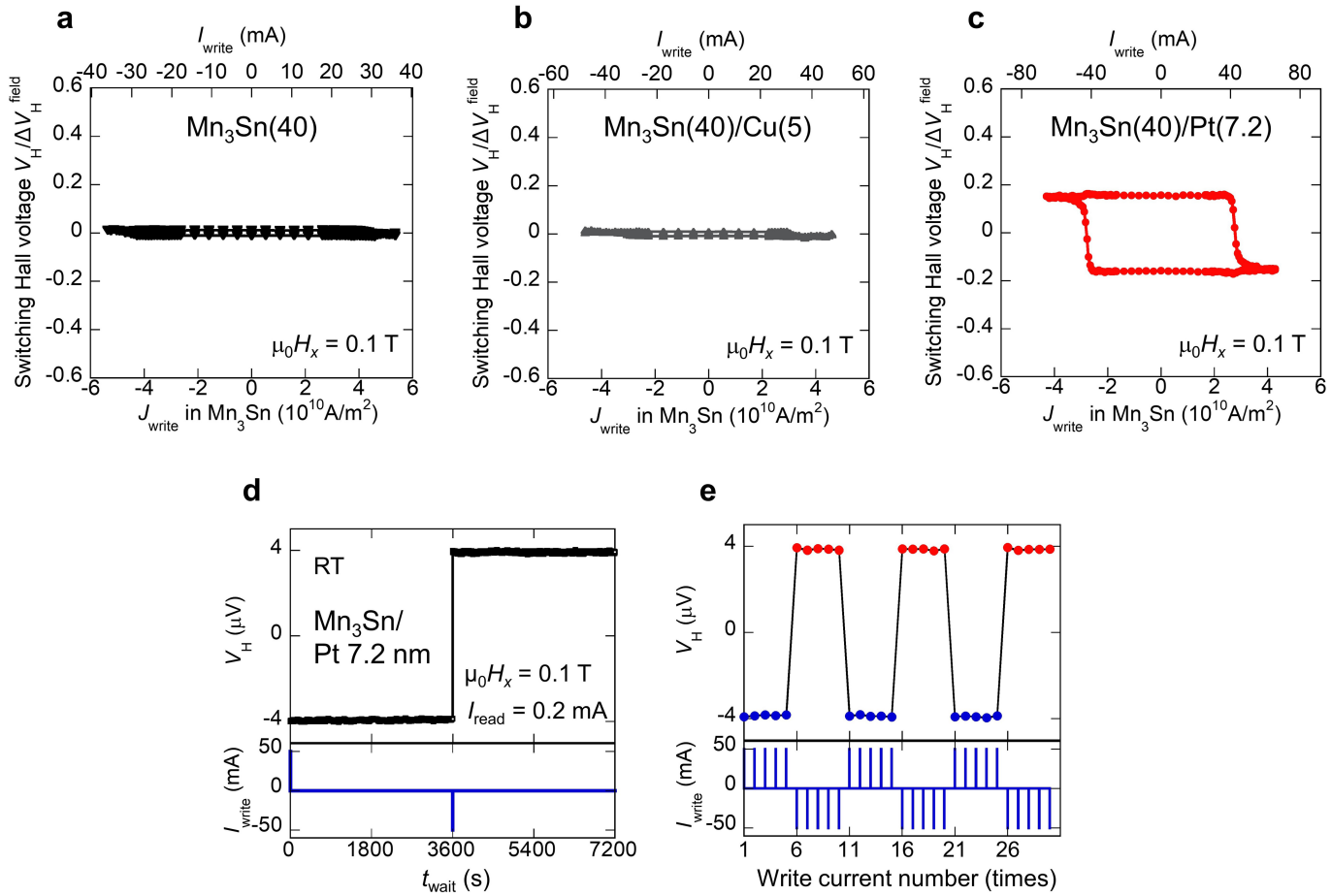
Extended Data Fig. 4 | Field-induced sign change in transverse thermoelectric conductivity and Hall conductivity of the Mn_3Sn thin film. **a, b**, Field dependence of the anomalous Nernst coefficient S_{ANE} and the Hall resistivity ρ_H (**a**) and the transverse thermoelectric conductivity α_N and Hall conductivity σ_H (**b**) of the the 40-nm-thick Mn_3Sn thin film at room temperature. The Seebeck coefficient S_{SE} and the resistivity ρ are also measured at 300 K and found to be constant (S_{SE} = 7.6 μV K^{-1} and ρ = 290 μΩ cm) in the field sweep measurements below ±4 T. α_N is estimated from the electrical conductivity $\sigma = 1/\rho$, the Hall conductivity $\sigma_H = -\rho_H/\rho^2$, the Seebeck coefficient S_{SE} and the Nernst coefficient S_{ANE} using the equation $\alpha_N = \sigma S_{ANE} + \sigma_H S_{SE}$ (Methods). Here, the heat current Q and electrical current I are applied parallel to the film plane and the field is applied along the normal direction (z direction) to the film plane.



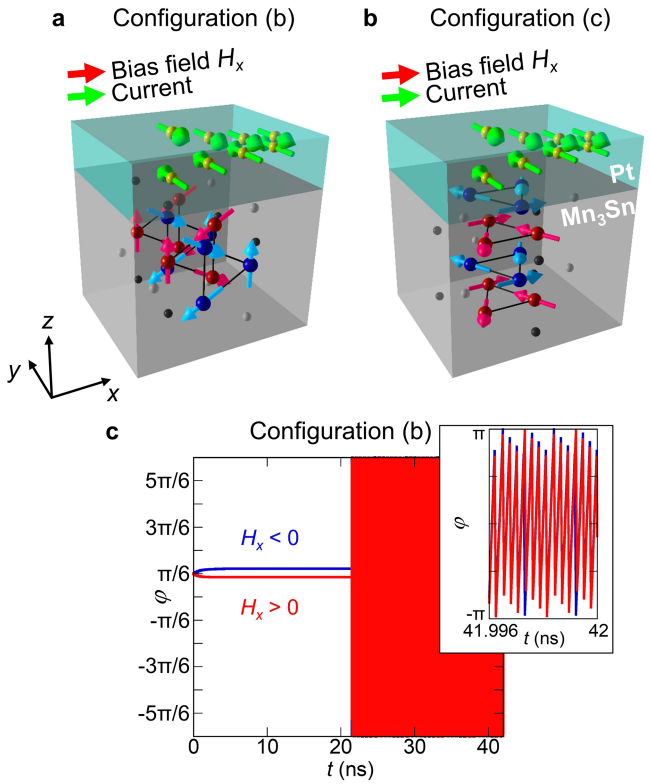
Extended Data Fig. 5 | Control of the nodal direction connecting a pair of Weyl points using the magnetic octupole polarization. a–**f**, Cluster magnetic octupole (orange arrow) consisting of the six spins on the kagome bilayer in real space (left) and schematic distributions of the Weyl points near the Fermi energy in momentum space (k_x–k_y plane at $k_z = 0$; right) for each magnetic structure of Mn_3Sn corresponding to $\varphi = \pi/6$ (**a**), $\pi/2$ (**b**), $5\pi/6$ (**c**), $-\pi/6$ (**d**), $-\pi/2$ (**e**) and $-5\pi/6$ (**f**). Red and blue spheres represent Weyl nodes that act as sources (+) and drains (−), respectively, of the Berry curvature (green arrows)[12].



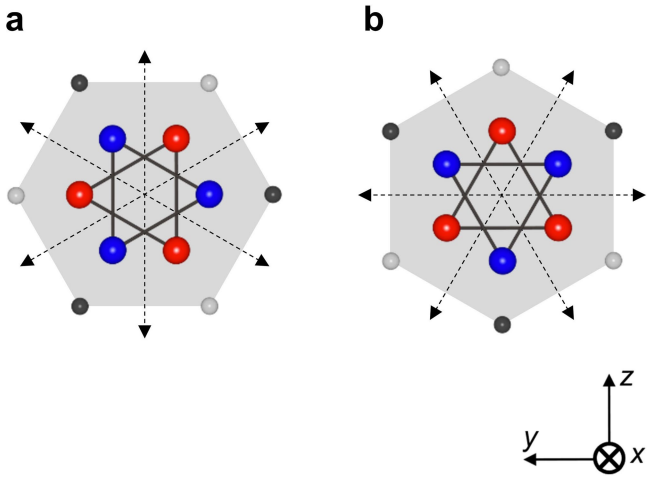
Extended Data Fig. 6 | Experimental conditions for electrical measurements. a, Sequence used for the SOT-induced switching measurements. **b**, Thickness dependence of the resistivity of the NM (Pt or W) layer obtained in the Si/SiO_2/Ru(2)/Mn_3Sn(40)/Pt or W(d_{NM})/AlO_x(5) Hall bar devices at room temperature.



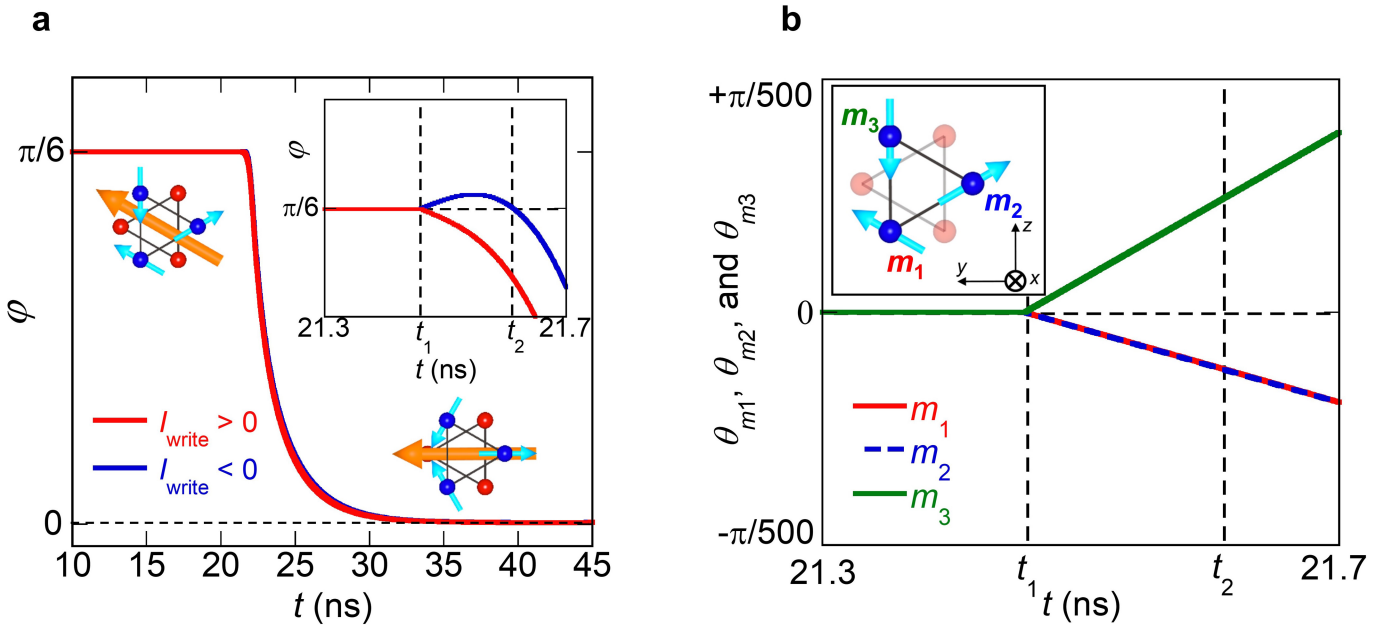
Extended Data Fig. 7 | Current-induced switching, signal stability and heating effects in the Mn_3Sn devices. a–**c**, Hall voltage versus write current density for Mn_3Sn without an NM layer (**a**), Mn_3Sn/Cu (**b**) and Mn_3Sn/Pt (**c**) Hall bar devices under a bias field of $H_x = 0.1$ T. In contrast to the Mn_3Sn/Pt device (**c**), which shows clear switching, the Hall voltage of the Mn_3Sn sample in **a** is not switched by the electric current, similarly to the Mn_3Sn/Cu sample in **b**. The top and bottom horizontal axes present the write current I_{write} in whole multilayers and the write current density J_{write} in the Mn_3Sn layer, respectively. **d**, Dependence of the Hall voltage V_H on the wait time (t_{wait}) measured after electrical switching of the AHE by the write current I_{write} (±50 mA, 100 ms) in the Mn_3Sn/Pt(7.2) device at room temperature. No variation of the AHE signal is observed for $t_{wait} = 600$ ms ≈ 1 h, which indicates that 600 ms is long enough to cool the sample down to room temperature, and the AHE signal is very stable in the Mn_3Sn Hall bar devices after the electrical switching. **e**, Hall voltage as a function of the number of write current pulses in the Mn_3Sn/Pt(7.2) device at room temperature. The AHE signal obtained after the first write current (±50 mA, 100 ms) does not change even after five consecutive pulses, similarly to FMs[29].



Extended Data Fig. 8 | Crystal grain configurations in the polycrystalline Mn_3Sn layer. a–**c**, Configurations of the SOT-induced switching in the Si/SiO_2/Ru(2)/Mn_3Sn(40)/Pt and W (d_{NM})/AlO_x(5) Hall bar devices. **a**, Configuration (b): the kagome layer is parallel to *I* and perpendicular to **p**. **b**, Configuration (c): the kagome layeris parallel to the current *I* and parallel to the electrically injected carrier spin polarization **p**. Green arrows represent the spin polarized current in the NM (for example, Pt) induced by the write current along the *x* direction. The crystal and magnetic structures of Mn_3Sn are presented (Fig. 1a, b). **c**, In-plane angle φ (as defined in Fig. 4c) of the octupole moment as a function of time *t* for the configuration (b). The coordinates *x*, *y* and *z* are defined as in **a**. The damping-like torque is applied at $t > 21.5$ ns. Inset, magnified view of the in-plane angle φ of the octupole moment as a function of time. The results indicate continuous rotation of a slightly canted ITS structure with a frequency of ~3.5 THz.



Extended Data Fig. 9 | Kagome plane arrangements for configuration (a). **a, b**, Kagome plane orientations (a-1) *b* axis [01$\bar{1}$0] parallel to *y* (**a**); (a-2) *a* axis [2$\bar{1}\bar{1}$0] parallel to *y* where the kagome layer is normal to the current *I* and parallel to *y*. The broken arrows represent the magnetic easy axis of the octupole polarization.



Extended Data Fig. 10 | Simulated dynamics of the sublattice moments during the switching. a, In-plane motions of the octupole polarization in the absence of a bias field. The red (blue) line corresponds to the motion under $I_{write} > 0$ ($I_{write} < 0$). Here, we use a write current with a finite rise time to suppress the incoherent oscillating behaviour. The inset shows a magnified view of the short period right after the current injection at t_1. The parameters used here are the same as those used in Fig. 4b. **b**, Evolution of the out-of-(kagome)plane components (parallel to the x direction) of the sublattice magnetic moments $\mathbf{m}_1$ (red), $\mathbf{m}_2$ (blue) and $\mathbf{m}_3$ (green), induced by $I_{write} < 0$. $\theta_{ma} > 0$ ($\theta_{ma} < 0$) ($a = 1, 2, 3$) corresponds to the positive (negative) component in the x direction from the y–z (kagome) plane.

Two-dimensional halide perovskite lateral epitaxial heterostructures

https://doi.org/10.1038/s41586-020-2219-7

Received: 14 August 2019

Accepted: 10 February 2020

Published online: 29 April 2020

Check for updates

Enzheng Shi[1,9], Biao Yuan[2,9], Stephen B. Shiring[1], Yao Gao[1], Akriti[1], Yunfan Guo[3], Cong Su[4], Minliang Lai[5], Peidong Yang[5,6,7], Jing Kong[3], Brett M. Savoie[1]✉, Yi Yu[2]✉ & Letian Dou[1,8]✉

Epitaxial heterostructures based on oxide perovskites and III–V, II–VI and transition metal dichalcogenide semiconductors form the foundation of modern electronics and optoelectronics[1–7]. Halide perovskites—an emerging family of tunable semiconductors with desirable properties—are attractive for applications such as solution-processed solar cells, light-emitting diodes, detectors and lasers[8–15]. Their inherently soft crystal lattice allows greater tolerance to lattice mismatch, making them promising for heterostructure formation and semiconductor integration[16,17]. Atomically sharp epitaxial interfaces are necessary to improve performance and for device miniaturization. However, epitaxial growth of atomically sharp heterostructures of halide perovskites has not yet been achieved, owing to their high intrinsic ion mobility, which leads to interdiffusion and large junction widths[18–21], and owing to their poor chemical stability, which leads to decomposition of prior layers during the fabrication of subsequent layers. Therefore, understanding the origins of this instability and identifying effective approaches to suppress ion diffusion are of great importance[22–26]. Here we report an effective strategy to substantially inhibit in-plane ion diffusion in two-dimensional halide perovskites by incorporating rigid π-conjugated organic ligands. We demonstrate highly stable and tunable lateral epitaxial heterostructures, multiheterostructures and superlattices. Near-atomically sharp interfaces and epitaxial growth are revealed by low-dose aberration-corrected high-resolution transmission electron microscopy. Molecular dynamics simulations confirm the reduced heterostructure disorder and larger vacancy formation energies of the two-dimensional perovskites in the presence of conjugated ligands. These findings provide insights into the immobilization and stabilization of halide perovskite semiconductors and demonstrate a materials platform for complex and molecularly thin superlattices, devices and integrated circuits.

Two-dimensional (2D) halide perovskites exhibit high photoluminescence quantum yield, long carrier lifetime and diffusion length, and remarkable optoelectronic tunability, owing to their structural and compositional flexibility[27–34]. These 2D perovskites form quantum wells composed of periodically repeating organic and inorganic layers along the out-of-plane direction, providing further structure and property tunability[35,36]. In this work, the synthesis of 2D halide perovskite lateral heterostructures is performed via a solution-phase sequential growth approach, as shown in Supplementary Fig. 1. In general, halide perovskites are susceptible to damage after two or more sequential growth steps, particularly when subsequent growth is performed in more aggressive conditions than the prior step—for example, under a higher temperature or using a more polar solvent. To eliminate the possibility of damaging the existing crystals, subsequent growth is performed under relatively milder growth conditions, such as lowering the growth temperature or adding more antisolvent in the precursor solution. Thus, subsequent halide perovskites are nucleated along the edges of the prior 2D crystals, thereafter forming concentric square/rectangular 2D lateral heterostructures directly on the SiO_2/Si substrate. By controlling the solution concentration and the growth temperature and time, the lateral size and vertical thickness of the 2D crystal can be controlled (Supplementary Figs. 2–4).

Lateral heterostructure formation

As shown in Fig. 1a, b, we create two types of lateral 2D halide perovskite heterostructures using different organic ligands: a conjugated ligand based on bithiophenylethylammonium ($2T^+$; Fig. 1a and Supplementary

[1]Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, USA. [2]School of Physical Science and Technology, ShanghaiTech University, Shanghai, China. [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Research Lab of Electronics (RLE), Massachusetts Institutes of Technology, Cambridge, MA, USA. [5]Department of Chemistry, University of California, Berkeley, CA, USA. [6]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [7]Kavli Energy NanoScience Institute, Berkeley, CA, USA. [8]Birck Nanotechnology Center, Purdue University, West Lafayette, IN, USA. [9]These authors contributed equally: Enzheng Shi, Biao Yuan. ✉e-mail: bsavoie@purdue.edu; yuyi1@shanghaitech.edu.cn; dou10@purdue.edu

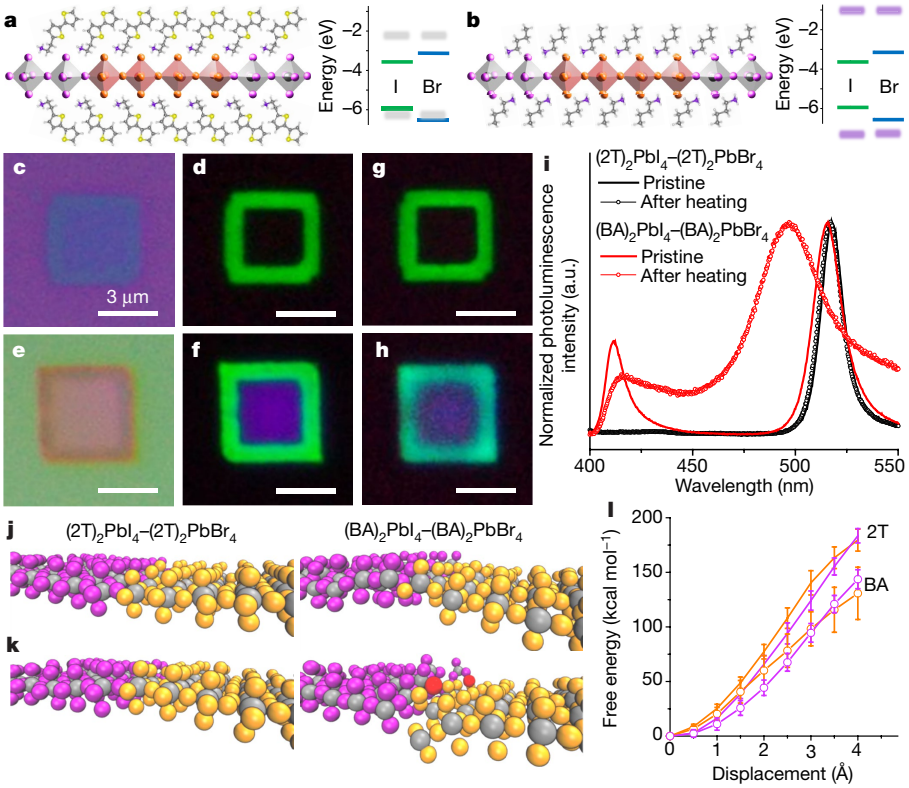


Fig. 1 | 2D halide perovskite lateral heterostructures stabilized by conjugated ligands. a, **b**, Schematic illustrations and proposed band alignments of $(2T)_2PbI_4$–$(2T)_2PbBr_4$ (**a**) and $(BA)_2PbI_4$–$(BA)_2PbBr_4$ (**b**) lateral heterostructures. The pairs of blue and green lines in the band diagrams at right represent the conduction band minimum and valence band maximum of inorganic $[PbBr_4]^{2-}$ and $[PbI_4]^{2-}$ octahedral layers, respectively. The broad, semi-transparent pairs of grey and purple lines correspond to the highest occupied molecular orbital and lowest unoccupied molecular orbital levels of the $2T^+$ and BA^+ organic layers, respectively. **c**, **d**, Optical (**c**) and photoluminescence (**d**) images of a $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructure. **e**, **f**, Optical (**e**) and photoluminescence (**f**) images of a $(BA)_2PbI_4$–$(BA)_2PbBr_4$ lateral heterostructure. **g**, Photoluminescence image of the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructure after 1 h of heating at 100 °C. **h**, Photoluminescence image of the $(BA)_2PbI_4$–$(BA)_2PbBr_4$ lateral heterostructure after 1 h of heating at 100 °C. All scale bars are 3 μm. **i**, Corresponding photoluminescence spectra of the heterostructures before and after heating. **j**, **k**, Snapshots from the molecular dynamics simulations at 298 K (**j**) and 800 K (**k**) for $(2T)_2PbI_4$–$(2T)_2PbBr_4$ (left) and $(BA)_2PbI_4$–$(BA)_2PbBr_4$ (right) showing the interface between each perovskite domain. For clarity, the organic ligands have been omitted. The colours correspond to: purple, iodine atoms; orange, bromine atoms; grey, lead atoms. In $(BA)_2PbI_4$–$(BA)_2PbBr_4$, the iodine atoms that have diffused across the interface and into the bromine domain are indicated in red. **l**, Free energy for removing a halide atom from an apical position to vacuum to generate a halide vacancy. The orange plots correspond to bromide perovskite and purple plots correspond to iodide perovskites. a.u., arbitrary units.

Table 1) and a widely studied alkyl ligand, butylammonium (BA^+; Fig. 1b). The heterostructures exhibit two concentric regions with slight optical contrast and distinct photoluminescence emissions, including the central square/rectangular bromide region and the peripheral iodide region with a uniform width of about 1–2 μm for the iodide region (Fig. 1c–f). For $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructures, no photoluminescence emission is detected from the central $(2T)_2PbBr_4$ region, which is attributed to type-II band alignment between the inorganic $[PbBr_4]^{2-}$ and organic ligand $2T^+$ layers as shown in the proposed band alignment (Fig. 1a)[37]. $(2T)_2PbI_4$ has a green photoluminescence emission (Fig. 1d, i) that peaks at 515 nm from the type-I band alignment between the $[PbI_4]^{2-}$ and $2T^+$ layers (Fig. 1a). The creation of the $(BA)_2PbI_4$–$(BA)_2PbBr_4$ heterostructure follows a similar synthetic procedure (Fig. 1e, f). It is known that the Br–I heterostructure is not thermally stable in three-dimensional perovskites because of the large solid-state diffusivity of Br^- and I^-. Interdiffusion of halide anions across the heterojunction can be triggered and accelerated upon mild heating[20,21]. Here, however, we found that when the conjugated 2T ligands are used, the interdiffusion between Br^- and I^- in the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructure is substantially inhibited. As shown in Fig. 1g, after heating at 100 °C for 1 h, no difference is observed in the photoluminescence image, and the photoluminescence emission shift is within 1 nm (Fig. 1i, with extended temperatures and times in Supplementary Fig. 5), suggesting negligible halide interdiffusion across the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ interface. By contrast, halide interdiffusion is fast in the reference heterostructure $(BA)_2PbI_4$–$(BA)_2PbBr_4$. As shown in Fig. 1h, i, after 1 h of heating at 100 °C, the interface between the blue region associated with $(BA)_2PbBr_4$ and the green region associated with $(BA)_2PbI_4$ becomes blurry, and there is a drastic change in the photoluminescence emission spectrum (extended temperatures and times are presented in Supplementary Fig. 6). The green peak initially located at 515 nm blueshifts to 496 nm, whereas the blue peak redshifts from 411 nm to 415 nm. Meanwhile, the peaks become much broader, indicating pronounced I/Br interdiffusion and alloying. It is estimated that the diffusion is slowed down by two to three orders of magnitude in the 2T heterostructures compared to the BA-based heterostructures (Supplementary Fig. 7).

Stabilization mechanism

Molecular dynamics simulations were used to investigate the suppression of halide interdiffusion in the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructure. The simulations on heterostructures of $(2T)_2PbI_4$–$(2T)_2PbBr_4$ and $(BA)_2PbI_4$–$(BA)_2PbBr_4$ at both room and elevated temperatures (298 K and 800 K, respectively) are consistent with the experimental observations: for $(2T)_2PbI_4$–$(2T)_2PbBr_4$, no diffusion is observed at

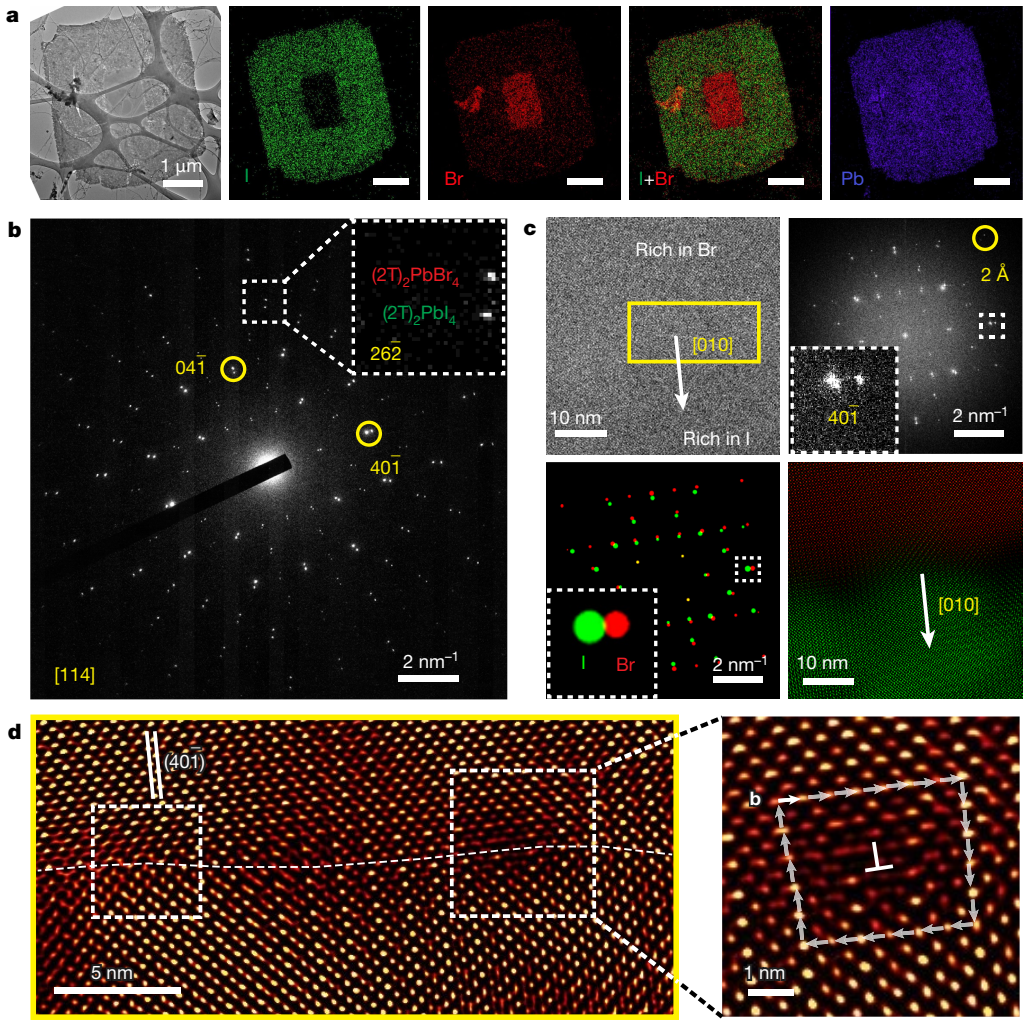


Fig. 2 | TEM characterization of the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructure. **a**, Low-magnification TEM image (left), followed by EDS elemental mappings of one $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructure, for (left to right): I, Br, I + Br and Pb. The scale bars are 1 μm. **b**, SAED pattern of the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ interface. Inset, enlarged view of one pair of splitting diffraction spots, which correspond to the $(26\bar{2})$ plane. The zone axis is [114]. The $(04\bar{1})$ and $(40\bar{1})$ planes are highlighted by yellow circles. **c**, AC-HRTEM and Fourier analysis. Top left, a selected AC-HRTEM image. Top right, Fourier transform pattern of the AC-HRTEM image. One pair of splitting spots is enlarged and shown in the inset, corresponding to the $(40\bar{1})$ plane in **b**. The outer spot at the top right is highlighted by a yellow circle; the corresponding lattice distance is 2 Å. Bottom left, false-colour Fourier masks for two sets of spots. One pair of masks for splitting spots is enlarged and shown in the inset. Bottom right, corresponding overlap (in false colour) of two sets of lattice fringes from $(2T)_2PbI_4$ (green) and $(2T)_2PbBr_4$ (red). **d**, Fourier filtered and magnified AC-HRTEM images of the yellow rectangular area in **c** (top left) showing the epitaxial interface. Left, the two solid white lines denote the $(40\bar{1})$ planes, and the two dashed white boxes highlight edge dislocations. Right, enlarged image of the rightmost highlighted edge dislocation; the bold white arrow denotes the Burgers vector **b**.

either temperature, whereas halide diffusion across the interface is observed in $(BA)_2PbI_4$–$(BA)_2PbBr_4$ at elevated temperatures. As shown visually in Fig. 1j, k, the amount of disorder at the interface is affected by the choice of organic ligand. In $(2T)_2PbI_4$–$(2T)_2PbBr_4$, the interface remains pristine and well ordered at elevated temperature. By contrast, interdiffusion and increased disorder are observed at the $(BA)_2PbI_4$–$(BA)_2PbBr_4$ heterojunction at elevated temperature. As comparative measures of the disorder, the displacements of the lead atoms from the 2D perovskite plane and octahedral tilt angles were summarized from the simulations (Supplementary Fig. 8). The displacements for $(2T)_2PbI_4$–$(2T)_2PbBr_4$ are less than around 2 Å at both room and elevated temperatures; by contrast, $(BA)_2PbI_4$-$(BA)_2PbBr_4$ shows displacements >3 Å and >4 Å at room temperature and elevated temperatures, respectively. Consistent trends are observed in the tilt angle results, with broader distributions occurring in proximity to the heterojunction (Supplementary Fig. 9). The increased disorder at room temperature is also consistent with the reduced resolution observed for BA-based compounds in the photoluminescence experiments (Supplementary Fig. 6). The larger and more rigid conjugated organic ligands are thus able to stabilize the interface and inorganic framework more effectively, whereas the smaller organic ligand leads to softer inorganic lattice that can accommodate the mismatch in halide sizes and facilitate halide interdiffusion.

Vacancy generation at high temperatures could also influence the interdiffusion process. For a vacancy-mediated diffusion process, the diffusion constant D can be expressed as[38]

$$D = D_0 \exp\left(-\frac{E_a}{k_B T}\right) = \alpha a^2 \omega N_v$$

$$N_v = \exp\left(-\frac{\Delta G}{k_B T}\right) = \exp\left(\frac{\Delta S}{k_B}\right)\exp\left(-\frac{\Delta H}{k_B T}\right)$$

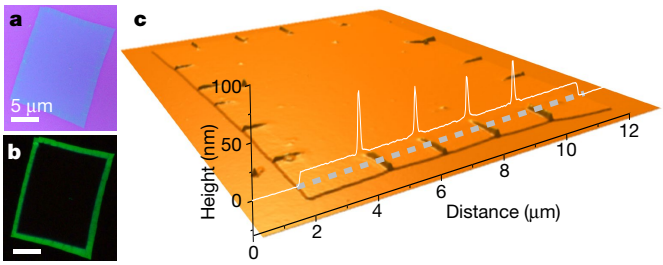


Fig. 3 | Periodic ripples in $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructures. **a**, **b**, Optical and photoluminescence images of a one-pot $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructure; ripples in the $(2T)_2PbI_4$ region can be distinguished. The scale bars are 5 μm. **c**, AFM image of the one-pot $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructure with clear ripples. The height profile along the grey dashed line indicates that the height of the ripples is around 40–50 nm.

where D_0 is the pre-exponential factor in the Arrhenius equation, E_a is the activation energy for the vacancy-assisted ion migration, N_v is the fraction of vacancies in the solid, α is a geometric constant, a is the elementary jump distance, ω is the jump frequency, k_B is the Boltzmann constant, T is the absolute temperature and ΔH, ΔS and ΔG are, respectively, the enthalpy change, the entropy change and the change in the Gibbs free energy for the formation of a vacancy. The vacancy concentration N_v is usually treated as a constant for most oxide materials and ω is the only temperature-dependent factor. However, for halide materials, the vacancy formation energy is normally much lower and therefore both N_v and ω are Arrhenius activated. Taking this into consideration, multistate Bennett acceptance ratio calculations on pure 2D perovskite systems were used to investigate halide and ligand vacancy formation (Fig. 1l). Using the BA ligand, the activation barrier to generate halide vacancies is much lower than that of the 2T ligand. Meanwhile, the free energy for creating a ligand vacancy by removing a ligand molecule from the surface is similar for both BA and 2T for a given halide species (Supplementary Fig. 10), indicating that the vacancy concentrations for the BA and 2T ligands are similar and may not play a dominant role. From these calculations, the number of ligand vacancies is expected to be similar upon heating the crystals, whereas a greater concentration of halide vacancies is to be expected in $(BA)_2PbI_4$–$(BA)_2PbBr_4$ heterojunctions, which also promotes faster anion interdiffusion. Therefore, it is suggested that both the lower interfacial disorder and the lower halide vacancy concentration in 2T-based perovskites probably contribute to the inhibition of the halide interdiffusion in the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterojunctions. As suggested by the differences between the free energy calculations for ligand and halide vacancy generation, the differences in the interfacial stability are largely attributed to the identity of the ligands and not to the identity of the individual halides.

Interface imaging and analysis

Energy dispersive X-ray spectroscopy (EDS) and selected area electron diffraction (SAED) were used to provide structural insights into the ultrathin $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructures. Because the heterostructures are directly grown on SiO_2/Si substrates, it is challenging to transfer the heterostructures to the TEM grids. To address this issue, we used a sacrificial polymethyl methacrylate (PMMA) layer on the growth substrate before the growth of heterostructures. The growth substrate with the PMMA layer was then soaked in chlorobenzene (see Methods) to release the 2D crystals, facilitating easy transfer of the 2D crystals to the TEM grid. The low-magnification TEM image and EDS elemental mappings of a $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructure (Fig. 2a) show that Pb is distributed uniformly throughout the heterostructure, whereas I and Br are concentrated in the peripheral and central regions, respectively. This is consistent with photoluminescence, X-ray diffraction (XRD) and scanning electron microscopy (SEM) characterizations (Fig. 1 and Supplementary Fig. 11). Hybrid perovskites are known to undergo rapid degradation under electron-beam irradiation. To avoid structural damage from the electron beams, the SAED patterns were obtained with a reduced electron dose rate (that is, less than 0.1 e Å^{-2} s^{-1}; e, electron). By comparing the experimental diffraction patterns with the kinematical simulated patterns, the zone axis for electron diffraction is determined to be [114] (see Supplementary Fig. 12 and related discussion). As shown in Supplementary Figs. 12, 13, the slight difference in diffraction patterns between pure $(2T)_2PbI_4$ and $(2T)_2PbBr_4$ sheets is attributed to their lattice mismatch (about 5–6%, Supplementary Tables 2, 3). For the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ interface, the SAED pattern (Fig. 2b) is composed of two sets of patterns with identical orientation from $(2T)_2PbI_4$ and $(2T)_2PbBr_4$, respectively, suggesting the epitaxial growth of $(2T)_2PbI_4$ from the $(2T)_2PbBr_4$ sheet.

To achieve a higher spatial resolution of the heterostructure interface while minimizing the radiation damage, low-dose aberration-corrected high-resolution transmission electron microscopy (AC-HRTEM) was performed at an accelerating voltage of 80 kV (Fig. 2c, d and Supplementary Fig. 14). In our early work, atomic-resolution imaging was achieved on all inorganic halide perovskites, taking advantage of low-dose AC-HRTEM[39]. Here, further incorporated with a minimum-dose strategy (Supplementary Fig. 14), the lattice information of radiation-sensitive hybrid organic–inorganic halide perovskites was successfully revealed. The continuous lattice fringes can be imaged at the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ interface (Fig. 2c, top left), and its corresponding Fourier transform information (Fig. 2c, top right) is consistent with the SAED pattern (Fig. 2b), further confirming the epitaxial growth between $(2T)_2PbBr_4$ and $(2T)_2PbI_4$ at the nanometre scale. A Fourier filtering technique was also used to further resolve the interface. Owing to the different lattice constants between the two segments across the interface, the real-space lattice information is obtained by applying well tuned masks in Fourier space. The masked green and red areas in Fig. 2c (bottom left) correspond to the lattice information from $(2T)_2PbI_4$ (Supplementary Fig. 15b) and $(2T)_2PbBr_4$ (Supplementary Fig. 15c), respectively. By superimposing the two inverse Fourier transform images, the near-atomically sharp interface can be better distinguished (Fig. 2c, bottom right). The clear interface is also observed in the strain mapping in Supplementary Fig. 16. Geometric phase images demonstrate the near-uniform lattice of the $(2T)_2PbBr_4$ layer and periodic lattice deformation of the epitaxial $(2T)_2PbI_4$ layer (Supplementary Fig. 16h). To relax the accumulated interfacial strain and stabilize the heteroepitaxy, periodic interfacial misfit dislocations are expected[40]. This is evidenced by our direct observations from the Fourier filtered AC-HRTEM images (Fig. 2d and Supplementary Fig. 17). Figure 2d clearly shows a misfit edge dislocation at the atomic level. In addition to the continuous lattice fringe along the $(40\bar{1})$ plane on both sides of the interface, the edge dislocations appear at around 15 nm intervals (Supplementary Fig. 17c). Furthermore, by analysing multiple samples, it was found that the $(2T)_2PbI_4$ preferentially grew along two directions: [100] and [010] (Supplementary Fig. 18). These observations suggest that the stable heteroepitaxy (Supplementary Fig. 19) between two similar perovskite structures is in-plane connected along the [100] and [010] directions with interfacial strain relaxed by misfit dislocations. The out-of-plane strain is probably relaxed through rearrangement of the organic cations.

In addition to forming misfit dislocations, strain can also be relaxed through ripple formation[7]. Here, we show that such ripples can be observed in the halide perovskite heterostructures. By simply mixing the precursors of $(2T)_2PbI_4$ and $(2T)_2PbBr_4$, one-pot synthesis of $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructures is achieved with a slightly less sharp interface (Fig. 3). The spontaneous formation of $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructures via the one-pot route is attributed to the large solubility difference between these two compounds (which leads to sequential precipitation from the solution during the solvent evaporation process) and the reduced nucleation

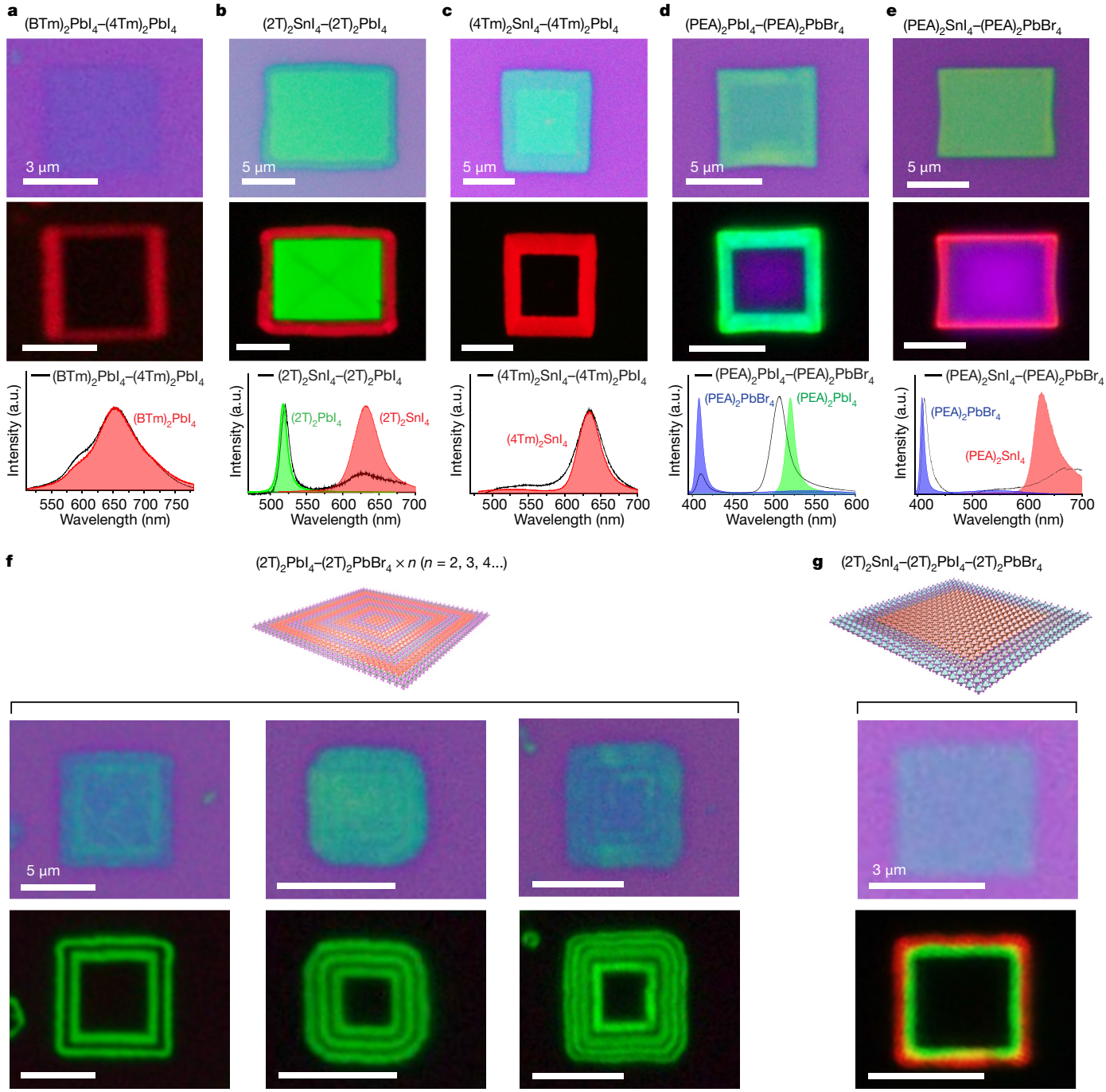


Fig. 4 | The library of 2D halide perovskite lateral heterostructures, multiheterostructures and superlattices. a–**e**, Lateral heterostructures of $(BTm)_2PbI_4$–$(4Tm)_2PbI_4$ (**a**), $(2T)_2SnI_4$–$(2T)_2PbI_4$ (**b**), $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ (**c**), $(PEA)_2PbI_4$–$(PEA)_2PbBr_4$ (**d**) and $(PEA)_2SnI_4$–$(PEA)_2PbBr_4$ (**e**). Top, optical images; middle, photoluminescence images; bottom, photoluminescence spectra. **f**, **g**, Schematic illustrations of the $(2T)_2PbI_4$–$(2T)_2PbBr_4 \times n$ lateral superlattice (**f**; left to right, $n = 2$, $n = 3$, $n = 4$) and $(2T)_2SnI_4$–$(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral multiheterostructure (**g**). Top images are optical and bottom images are photoluminescence. Scale bars for **a** and **g** are 3 μm, all other scale bars are 5 μm.

energy of $(2T)_2PbI_4$ along the edge of $(2T)_2PbBr_4$ (see Supplementary Figs. 20, 21 and related discussions). In the $(2T)_2PbI_4$ regions in these heterostructures, periodic ripples formed—an indication of a coherent interface free of dislocations. From the atomic force microscopy (AFM) image of the heterostructure, the height of the ripples was found to be around 40–50 nm and the inter-distance between adjacent ripples is about 1.6 μm, which agrees with the 5–6% lattice strain. These ripple structures are larger and less dense than those observed in WS_2–WSe_2 heterostructures (about 1–2 nm high, approximately every 30 nm)[7]; it is surprising that the weak and soft halide perovskite lattice is able to preserve the coherency at the heterointerface so well. Lateral interface engineering—including the interfacial dislocations and ripples discovered in 2D halide perovskite heterostructures—provides insights into and opportunities to better control the interface integrity and associated electronic and optoelectronic properties.

Multicomplex heterostructure formation

We further demonstrate the general synthesis of various lateral heterostructures (between different halides, metal cations and

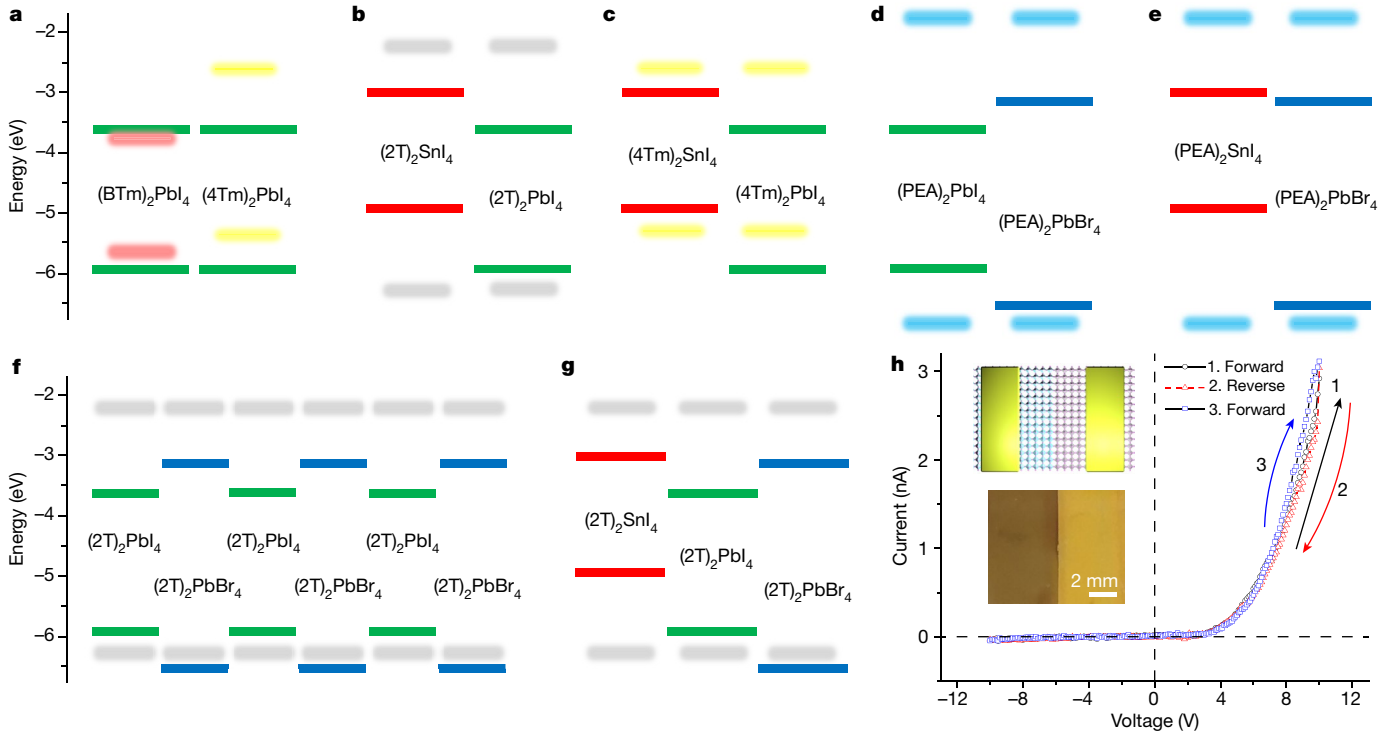


Fig. 5 | Proposed band alignments and electrical characteristics of the heterostructures. a–**g**, Proposed band alignments for $(BTm)_2PbI_4$–$(4Tm)_2PbI_4$ (**a**), $(2T)_2SnI_4$–$(2T)_2PbI_4$ (**b**), $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ (**c**), $(PEA)_2PbI_4$–$(PEA)_2PbBr_4$ (**d**), $(PEA)_2SnI_4$–$(PEA)_2PbBr_4$ (**e**), $(2T)_2PbI_4$–$(2T)_2PbBr_4 \times 3$ superlattice (**f**) and $(2T)_2SnI_4$–$(2T)_2PbI_4$–$(2T)_2PbBr_4$ (**g**). The pairs of blue, green and red lines represent the conduction band minimums and valence band maximums of inorganic $[PbBr_4]^{2-}$, $[PbI_4]^{2-}$ and $[SnI_4]^{2-}$ octahedral layers, respectively. The pairs of broad, semi-transparent grey, red, yellow and blue lines correspond to the highest occupied molecular orbitals and lowest unoccupied molecular orbitals of the $2T^+$, BTm^+, $4Tm^+$ and PEA^+ organic layers, respectively. Band bending and Fermi level matching is not considered as these values have not yet been reliably measured and reported. **h**, Rectification behaviour of a $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ heterostructure diode device. The labels 1, 2, and 3 denote the current–voltage scan direction and order. Inset, schematic illustration and optical image of the thin-film $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ heterostructure.

organic ligands) and even superlattices of 2D halide perovskites following the solution-phase epitaxial growth strategy. Additional π-conjugated organic ligands, including phenylethylammonium (PEA^+), quaterthiophenylethylammonium ($4Tm^+$) and 7-(thiophen-2-yl) benzothiadiazol-4-yl)-[2,2'-bithiophen]-5-yl)ethylammonium (BTm^+), were used. The chemical structures of all ligands in this study are listed in Supplementary Fig. 22a. Corresponding optical and photoluminescence images of each type of 2D sheet are shown in Supplementary Fig. 22b–k. The XRD patterns and photoluminescence spectra of these 2D sheets are summarized in Supplementary Fig. 23. Using these ligands, we synthesized a $(BTm)_2PbI_4$–$(4Tm)_2PbI_4$ heterostructure between different organic ligands, $(2T)_2SnI_4$–$(2T)_2PbI_4$ and $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ heterostructures between different metal cations, a $(PEA)_2PbI_4$–$(PEA)_2PbBr_4$ heterostructure between different halides, a $(PEA)_2SnI_4$–$(PEA)_2PbBr_4$ heterostructure between different halides and metals, $(2T)_2PbI_4$–$(2T)_2PbBr_4 \times n$ ($n = 2, 3, 4$) superlattices and $(2T)_2SnI_4$–$(2T)_2PbI_4$–$(2T)_2PbBr_4$ multiheterostructures (Fig. 4).

As shown in Fig. 4a, lateral heterostructures can be created between two segments with the distinct organic ligands $4Tm^+$ and BTm^+ (we note that the sizes for two different organic ligands should be similar to reduce strain). The $(4Tm)_2PbI_4$ region shows no photoluminescence emission, owing to the type-II band alignment between the $4Tm^+$ ligand and the $[PbI_4]^{2-}$ layers (Supplementary Figs. 22f, 23)[41]. The broad red photoluminescence emission from the peripheral region of the heterostructure originates from $(BTm)_2PbI_4$, which is attributed to the type-I band alignment between the $[PbI_4]^{2-}$ layer and the BTm^+ layer (Supplementary Figs. 22g, 23). In addition, the bandgap and optoelectronic properties of the halide perovskite materials can be modulated by altering the metal atoms centring the $[MX_4]^{2-}$ octahedrons. By substituting Pb with Sn, 2D halide perovskite sheets were successfully synthesized with $2T^+$ and $4Tm^+$ ligands, including $(2T)_2SnI_4$ and $(4Tm)_2SnI_4$ (Supplementary Figs. 22h, i). In combination with the Pb-based perovskites, both $(2T)_2SnI_4$–$(2T)_2PbI_4$ and $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ lateral heterostructures were created (Fig. 4b, c). Two distinct photoluminescence emission peaks from the $(2T)_2SnI_4$–$(2T)_2PbI_4$ heterostructure correspond to pure $(2T)_2SnI_4$ and $(2T)_2PbI_4$, and the gap in the photoluminescence image between the Pb and Sn perovskite segments is probably induced by exciton dissociation at the interface. Similarly, Pb and Sn perovskite segments in the $(4Tm)_2PbI_4$–$(4Tm)_2SnI_4$ heterostructure exhibit a slight contrast in the optical image and distinct emission colours in the photoluminescence image (Fig. 4c). The peripheral region exhibits an identical photoluminescence spectrum with the reference $(4Tm)_2SnI_4$ crystal (Supplementary Fig. 23). Another, smaller, conjugated ligand, PEA^+, was used to construct a lateral heterostructure between Br and I components. As shown in Fig. 4d, the photoluminescence image and spectrum of the $(PEA)_2PbI_4$–$(PEA)_2PbBr_4$ heterostructure are close to those of $(BA)_2PbI_4$–$(BA)_2PbBr_4$, showing a purple–blue colour in the centre and green in the peripheral region. In addition, we have created $(PEA)_2SnI_4$–$(PEA)_2PbBr_4$ lateral heterostructures based on different halides and different metals at the same time, where the inorganic backbone is $[PbBr_4]^{2-}$ in the interior region and $[SnI_4]^{2-}$ in the peripheral region (Fig. 4e).

Apart from the two-segment concentric heterostructures, more complex heterostructures have also been demonstrated. As shown in Fig. 4f, the $(2T)_2PbI_4$–$(2T)_2PbBr_4 \times n$ ($n = 2, 3, 4$) superlattices are synthesized through multiple repeated growth steps. As indicated in the schematic illustration and the photoluminescence image, 4–8 concentric rectangles are formed, with green emission regions representing

$(2T)_2PbI_4$ and quenched regions representing $(2T)_2PbBr_4$. Additionally, multiheterostructures are realized by a third growth of the $(2T)_2SnI_4$ layer with a red emission along the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructure (Fig. 4g). Additionally, the synthetic yields of the above-mentioned heterostructures are very high, as illustrated in the lower magnification optical and photoluminescence images (Supplementary Fig. 24). The thickness of the heterostructures ranges from a single cell to a few unit cells, as demonstrated by the AFM study (Supplementary Fig. 25).

Discussion

These 2D heterostructures exhibit useful optical and electronic properties. For instance, as summarized in Fig. 5a–g[41–43], the band alignments of these heterostructures can be modulated either by varying the inorganic composition in the lateral in-plane direction or by modifying the molecular structure in the out-of-plane direction. To the best of our knowledge, such multicomplex integrated systems have not previously been realized in other nanoscale heterostructures. To investigate the electronic properties and demonstrate potential device applications of the heterostructures, we have fabricated a proof-of-concept thin-film electrical diode based on a $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ heterostructure with a type-II band alignment (Fig. 5h). Stable electrical rectifying behaviour with a rectification ratio of around 10^2 was observed without hysteresis. Furthermore, we observed an enhanced exciton lifetime at the interface of the $(2T)_2PbBr_4$–$(2T)_2PbI_4$ heterostructure and the $(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ heterostructure from the fluorescence-lifetime imaging measurements, which shed light upon the possibility of tuning the optoelectronic properties via lattice-strain engineering at the interfaces of the heterostructures and superlattices (Supplementary Fig. 26 and Supplementary Table 4).

We have shown that the ion interdiffusion across 2D halide perovskite heterojunctions can be substantially inhibited by using rigid conjugated ligands in the 2D perovskite structure. This suppression of ion migration enables stable and near-atomically sharp interfaces to be obtained via sequential epitaxial growth. The generic synthesis of a wide range of 2D lateral halide perovskite heterostructures, superlattices and multiheterostructures not only presents a powerful platform for advancing the fundamental crystal chemistry of halide perovskites, but also opens up the possibility of further exploring their optoelectronic properties and their applications in diodes, lasers, transistors and photovoltaic devices. Particularly, the role of the conjugated organic ligands on the electronic properties of the heterostructures is worthy of further investigation. More importantly, these findings and methodologies may be extended to other classes of solution-processed 2D nanomaterials.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2219-7.

1. Lugli, P. & Goodnick, S. M. Nonequilibrium longitudinal-optical phonon effects in GaAs–AlGaAs quantum wells. *Phys. Rev. Lett.* **59**, 716–719 (1987).
2. Ahn, C. H., Rabe, K. M. & Triscone, J.-M. Ferroelectricity at the nanoscale: local polarization in oxide thin films and heterostructures. *Science* **303**, 488–491 (2004).
3. Bernevig, B. A., Hughes, T. L. & Zhang, S.-C. Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science* **314**, 1757–1761 (2006).
4. Gong, Y. et al. Vertical and in-plane heterostructures from WS_2/MoS_2 monolayers. *Nat. Mater.* **13**, 1135–1142 (2014).
5. Zhang, Z. et al. Robust epitaxial growth of two-dimensional heterostructures, multiheterostructures, and superlattices. *Science* **357**, 788–792 (2017).
6. Sahoo, P. K., Memaran, S., Xin, Y., Balicas, L. & Gutiérrez, H. R. One-pot growth of two-dimensional lateral heterostructures via sequential edge-epitaxy. *Nature* **553**, 63–67 (2018).
7. Xie, S. et al. Coherent, atomically thin transition-metal dichalcogenide superlattices with engineered strain. *Science* **359**, 1131–1136 (2018).
8. Kojima, A., Teshima, K., Shirai, Y. & Miyasaka, T. Organometal halide perovskites as visible-light sensitizers for photovoltaic cells. *J. Am. Chem. Soc.* **131**, 6050–6051 (2009).
9. Tsai, H. et al. High-efficiency two-dimensional Ruddlesden–Popper perovskite solar cells. *Nature* **536**, 312–316 (2016).
10. Bai, S. et al. Planar perovskite solar cells with long-term stability using ionic liquid additives. *Nature* **571**, 245–250 (2019).
11. Cao, Y. et al. Perovskite light-emitting diodes based on spontaneously formed submicrometre-scale structures. *Nature* **562**, 249–253 (2018).
12. Yakunin, S. et al. Detection of X-ray photons by solution-processed lead halide perovskites. *Nat. Photon.* **9**, 444–449 (2015).
13. Wei, H. et al. Sensitive X-ray detectors made of methylammonium lead tribromide perovskite single crystals. *Nat. Photon.* **10**, 333–339 (2016).
14. Feng, J. et al. Single-crystalline layered metal-halide perovskite nanowires for ultrasensitive photodetectors. *Nat. Electron.* **1**, 404–410 (2018).
15. Sutherland, B. R. & Sargent, E. H. Perovskite photonic sources. *Nat. Photon.* **10**, 295–302 (2016).
16. Snaith, H. J. Present status and future prospects of perovskite photovoltaics. *Nat. Mater.* **17**, 372–376 (2018).
17. Berry, J. et al. Hybrid organic–inorganic perovskites (HOIPs): opportunities and challenges. *Adv. Mater.* **27**, 5102–5112 (2015).
18. Akkerman, Q. A. et al. Tuning the optical properties of cesium lead halide perovskite nanocrystals by anion exchange reactions. *J. Am. Chem. Soc.* **137**, 10276–10281 (2015).
19. Hoffman, J. B., Lennart Schleper, A. & Kamat, P. V. Transformation of sintered $CsPbBr_3$ nanocrystals to cubic $CsPbI_3$ and gradient $CsPbBr_xI_{3-x}$ through halide exchange. *J. Am. Chem. Soc.* **138**, 8603–8611 (2016).
20. Lai, M. et al. Intrinsic anion diffusivity in lead halide perovskites is facilitated by a soft lattice. *Proc. Natl Acad. Sci. USA* **115**, 11929–11934 (2018).
21. Pan, D. et al. Visualization and studies of ion-diffusion kinetics in cesium lead bromide perovskite nanowires. *Nano Lett.* **18**, 1807–1813 (2018).
22. Park, N.-G., Grätzel, M., Miyasaka, T., Zhu, K. & Emery, K. Towards stable and commercially available perovskite solar cells. *Nat. Energy* **1**, 16152 (2016).
23. Rong, Y. et al. Challenges for commercializing perovskite solar cells. *Science* **361**, eaat8235 (2018).
24. Park, B. & Seok, S. I. Intrinsic instability of inorganic–organic hybrid halide perovskite materials. *Adv. Mater.* **31**, 1805337 (2019).
25. Wang, Z. et al. Efficient ambient-air-stable solar cells with 2D–3D heterostructured butylammonium-caesium-formamidinium lead halide perovskites. *Nat. Energy* **2**, 17135 (2017).
26. Huang, Z. et al. Suppressed ion migration in reduced-dimensional perovskites improves operating stability. *ACS Energy Lett.* **4**, 1521–1527 (2019).
27. Dou, L. et al. Atomically thin two-dimensional organic–inorganic hybrid perovskites. *Science* **349**, 1518–1521 (2015).
28. Jemli, K. et al. Two-dimensional perovskite activation with an organic luminophore. *ACS Appl. Mater. Interfaces* **7**, 21763–21769 (2015).
29. Connor, B. A., Leppert, L., Smith, M. D., Neaton, J. B. & Karunadasa, H. I. Layered halide double perovskites: dimensional reduction of $Cs_2AgBiBr_6$. *J. Am. Chem. Soc.* **140**, 5235–5240 (2018).
30. Leng, K. et al. Molecularly thin two-dimensional hybrid perovskites with tunable optoelectronic properties due to reversible surface relaxation. *Nat. Mater.* **17**, 908–914 (2018).
31. Zhang, Q., Chu, L., Zhou, F., Ji, W. & Eda, G. Excitonic properties of chemically synthesized 2D organic–inorganic hybrid perovskite nanosheets. *Adv. Mater.* **30**, 1704055 (2018).
32. Spanopoulos, I. et al. Uniaxial expansion of the 2D Ruddlesden–Popper perovskite family for improved environmental stability. *J. Am. Chem. Soc.* **141**, 5518–5534 (2019).
33. Cortecchia, D. et al. Broadband emission in two-dimensional hybrid perovskites: the role of structural deformation. *J. Am. Chem. Soc.* **139**, 39–42 (2017).
34. Yang, S. et al. Ultrathin two-dimensional organic–inorganic hybrid perovskite nanosheets with bright, tunable photoluminescence and high stability. *Angew. Chem. Int. Ed.* **56**, 4252–4255 (2017).
35. Saparov, B. & Mitzi, D. B. Organic–inorganic perovskites: structural versatility for functional materials design. *Chem. Rev.* **116**, 4558–4596 (2016).
36. Ortiz-Cervantes, C. et al. Thousand-fold conductivity increase in 2D perovskites by polydiacetylene incorporation and doping. *Angew. Chem. Int. Ed.* **57**, 13882–13886 (2018).
37. Liu, C. et al. Tunable semiconductors: control over carrier states and excitations in layered hybrid organic–inorganic perovskites. *Phys. Rev. Lett.* **121**, 146401 (2018).
38. Borg, R. J. & Dienes, G. J. *An Introduction to Solid State Diffusion* (Academic Press, 1988).
39. Yu, Y. et al. Atomic resolution imaging of halide perovskites. *Nano Lett.* **16**, 7530–7535 (2016).
40. Matthews, J. Defects in epitaxial multilayers I. Misfit dislocations. *J. Cryst. Growth* **27**, 118–125 (1974).
41. Gao, Y. et al. Molecular engineering of organic–inorganic hybrid perovskites quantum wells. *Nat. Chem.* **11**, 1151–1157 (2019).
42. Gao, Y. et al. Highly stable lead-free perovskite field-effect transistors incorporating linear π-conjugated organic ligands. *J. Am. Chem. Soc.* **141**, 15577–15585 (2019).
43. Silver, S., Yin, J., Li, H., Brédas, J. L. & Kahn, A. Characterization of the valence and conduction band levels of $n = 1$ 2D perovskites: a combined experimental and theoretical investigation. *Adv. Energy Mater.* **8**, 1703468 (2018).

Methods

Solution-phase synthesis of pure 2D halide perovskite sheets

In this study, ten types of pure 2D halide perovskite sheets were synthesized via a quaternary solvent method.

Chemicals and reagents. Organic solvents, including anhydrous chlorobenzene (CB), dimethylformide (DMF), acetonitrile (AN) and dichlorobenzene (DCB) and solid chemicals, including lead bromide ($PbBr_2$), lead iodide (PbI_2) and tin iodide (SnI_2), were purchased commercially (Sigma Aldrich). *n*-butylammonium bromide (BA·HBr), *n*-butylammonium iodide (BA·HI), phenethylammonium bromide (PEA·HBr) and phenethylammonium iodide (PEA·HI) were purchased commercially (Greatcell Solar). All above chemicals were used as received. Other ammonium salt ligands, including 2T·HI, 2T·HBr, 4Tm·HI and BTm·HI were synthesized in our lab[41].

Synthesis of 2D halide perovskite sheets. 10 μmol of MX_2 (M = Pb or Sn, X = Br or I) and 20 μmol of L·HX (L = BA, PEA, 2T, 4Tm or BTm, X = Br or I) were dissolved in 2 ml of DMF/CB co-solvent (1:1 volume ratio) to prepare 5 mM of stock solution. The stock solution was then diluted 120 times by CB/AN/DCB co-solvent (2.5:1:0.01 volume ratio). 5–10 μl of diluted solution was added onto the growth substrate SiO_2 (300 nm)/Si, which was placed at the bottom of a 4 ml glass vial. The 4 ml vial was then transferred into a secondary glass vial (20 ml) containing 3 ml of CB. After that, the secondary vial was capped and moved onto a 70 °C hot plate. The antisolvent (CB) inside the secondary vial slows down the solvent evaporation from the substrate and facilitates the formation of large 2D halide perovskite sheets. The growth typically took 10–30 min. Solution preparation and sheets growth were carried out in a N_2-filled glovebox. The growth substrate was cleaned in piranha acid for 2 h before use.

The growth steps for all the halide perovskite sheets in Supplementary Fig. 22 were the same. However, the growth parameters for some sheets were slightly different. The above recipe applies to the growth of $(2T)_2PbBr_4$, $(BA)_2PbBr_4$, $(BA)_2PbI_4$ and $(PEA)_2PbBr_4$. For $(2T)_2PbI_4$, $(2T)_2SnI_4$ and $(PEA)_2PbI_4$, the stock solution was diluted 120 times by CB/AN/DCB co-solvent (9.5:1:0.01 volume ratio). For $(4Tm)_2PbI_4$ and $(4Tm)_2SnI_4$, the stock solution was diluted 240 times by CB/AN/DCB co-solvent (3.9:1:0.01 volume ratio) and the growth temperature was 90 °C. For $(BTm)_2PbI_4$, the stock solution was diluted 1,440 times by CB/AN/DCB co-solvent (7.4:1:0.01 volume ratio) and the growth temperature was 90 °C.

For the quaternary solvent method for crystal growth, CB helps reduce the solubility of 2D perovskite in DMF and promotes the sheet crystallization. AN has a lower boiling point compared to CB, and 2D perovskite has a limited solubility in AN. In this case, AN evaporates more quickly and initiates the 2D perovskite nucleation at a relatively low concentration, thus decreasing the thickness of sheets. Moreover, compared to CB, DCB has a higher boiling point. The addition of DCB avoids the gradually increasing concentration of perovskite solution with the evaporation of AN and CB, thus ensuring the uniform distribution of sheets throughout the growth substrate.

Epitaxial synthesis of 2D lateral halide perovskite heterostructures

The epitaxial growth of 2D lateral halide perovskite heterostructures was based on the above quaternary solvent method. The growth of the first sheet is identical to above method. To eliminate the possibility of crystal damage, subsequent growth was performed under milder growth conditions in our study—that is, by lowering the growth temperature or adding more antisolvent in the solution.

$(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructures and superlattices. The sequence of the two steps for heterostructure formation (Br followed by I versus I followed by Br) is dictated by the solubility difference of the two halide perovskites in the solvent system. As $(2T)_2PbBr_4$ has a lower solubility in the quaternary solvent, it is synthesized before $(2T)_2PbI_4$. The $(2T)_2PbI_4$ stock solution was diluted 480 times by CB/AN/DCB co-solvent (6:1:0.01 volume ratio) for the growth of heterostructures. After the growth of $(2T)_2PbBr_4$ sheets, the hot plate was cooled down to 50 °C. Then 10 μl of the diluted $(2T)_2PbI_4$ solution was added onto the growth substrate. The epitaxial growth of $(2T)_2PbI_4$ along the $(2T)_2PbBr_4$ sheets typically took about 30 min. The $(2T)_2PbI_4$–$(2T)_2PbBr_4 \times n$ superlattices were realized by *n* repeated growths of $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructures.

$(BA)_2PbI_4$–$(BA)_2PbBr_4$ heterostructures. The $(BA)_2PbI_4$ stock solution was diluted 240 times by CB/AN/DCB co-solvent (6:1:0.01 volume ratio) for the growth of heterostructures. After the growth of the $(BA)_2PbBr_4$ sheets, the hot plate was cooled down to 50 °C. Then 10 μl of the diluted $(BA)_2PbI_4$ solution was added onto the growth substrate. The epitaxial growth of $(BA)_2PbI_4$ along $(BA)_2PbBr_4$ sheets typically took about 30 min.

$(BTm)_2PbI_4$–$(4Tm)_2PbI_4$ heterostructures. The $(BTm)_2PbI_4$ stock solution was diluted 1,440 times by CB/AN/DCB co-solvent (3.2:1:0.01 volume ratio) for the growth of heterostructures. After the growth of the $(4Tm)_2PbI_4$ sheets, the hot plate was cooled down to 50 °C. Then 10 μl of the diluted $(BTm)_2PbI_4$ solution was added onto the growth substrate. The epitaxial growth of $(BTm)_2PbI_4$ along $(4Tm)_2PbI_4$ sheets typically took about 30 min.

$(2T)_2SnI_4$–$(2T)_2PbI_4$ heterostructures. The $(2T)_2SnI_4$ stock solution was diluted 120 times by CB/AN/DCB co-solvent (9.5:1:0.01 volume ratio) for the growth of heterostructures. After the growth of the $(2T)_2PbI_4$ sheets, the hot plate was cooled down to 50 °C. Then 10 μl of the diluted $(2T)_2SnI_4$ solution was added onto the growth substrate. The epitaxial growth of $(2T)_2SnI_4$ along the $(2T)_2PbI_4$ sheets typically took about 30 min.

$(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ heterostructures. The $(4Tm)_2SnI_4$ stock solution was diluted 240 times by CB/AN/DCB co-solvent (1.8:1:0.01 volume ratio) for the growth of heterostructures. After the growth of the $(4Tm)_2PbI_4$ sheets, the hot plate was cooled down to 80 °C. Then 10 μl of the diluted $(4Tm)_2SnI_4$ solution was added onto the growth substrate. The epitaxial growth of $(4Tm)_2SnI_4$ along the $(4Tm)_2PbI_4$ sheets typically took about 10 min.

$(PEA)_2PbI_4$–$(PEA)_2PbBr_4$ heterostructures. The $(PEA)_2PbI_4$ stock solution was diluted 240 times by CB/AN/DCB co-solvent (9.5:1:0.01 volume ratio) for the growth of heterostructures. After the growth of the $(PEA)_2PbBr_4$ sheets, the hot plate was cooled down to 50 °C. Then 10 μl of the diluted $(PEA)_2PbI_4$ solution was added onto the growth substrate. The epitaxial growth of $(PEA)_2PbI_4$ along the $(PEA)_2PbBr_4$ sheets typically took about 30 min.

$(PEA)_2SnI_4$–$(PEA)_2PbBr_4$ heterostructures. The $(PEA)_2SnI_4$ stock solution was diluted 580 times by CB/AN/DCB co-solvent (35:2:0.01 volume ratio) for the growth of heterostructures. After the growth of the $(PEA)_2PbBr_4$ sheets, the hot plate was cooled down to 50 °C. Then 10 μl of the diluted $(PEA)_2SnI_4$ solution was added onto the growth substrate. The epitaxial growth of $(PEA)_2PbI_4$ along $(PEA)_2PbBr_4$ sheets typically took about 30 min.

$(2T)_2SnI_4$–$(2T)_2PbI_4$-$(2T)_2PbBr_4$ multiheterostructures. $(2T)_2SnI_4$ stock solution was diluted 480 times by CB/AN/DCB co-solvent (9.5:1:0.01 volume ratio) for the growth of multiheterostructures. Following the growth of the $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructures, 10 μl of the diluted $(2T)_2SnI_4$ solution was added onto the growth substrate. The epitaxial

growth of $(2T)_2SnI_4$ along the $(2T)_2PbI_4$ sheets typically took about 30 min. The growth temperature was set at 50 °C.

One-pot synthesis of $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructures. The $(2T)_2PbI_4$ and $(2T)_2PbBr_4$ stock solutions were mixed in a 1:1 ratio and diluted 480 times by CB/AN/DCB co-solvent (6:1:0.01 volume ratio) for the one-pot growth of heterostructures. 10 µl of the diluted solution was added onto the growth substrate at a temperature of 70 °C. The Br and I components phase-separated spontaneously during the evaporation of the solvents. The coherent epitaxial growth of $(2T)_2PbI_4$ along the $(2T)_2PbBr_4$ sheets typically took about 10 min.

$(4Tm)_2SnI_4$–$(4Tm)_2PbI_4$ thin film lateral heterostructure devices. $(4Tm)_2PbI_4$ (0.1 mol l^{-1}) was spin-coated on insulating substrates (quartz or SiO_2/Si), followed by 180 °C annealing on a hot plate for 5 min. Then, half of the $(4Tm)_2PbI_4$ film was removed by razor blade, and the rest of the film was covered by a Kapton tape. The sample was then treated by ultraviolet ozone for 30–60 min to make the exposed area hydrophilic. The Kapton tape was removed and $(4Tm)_2SnI_4$ (0.1 mol l^{-1}) was spin-coated on the ultraviolet ozone-treated sample, followed by 175 °C annealing for 5 min. Finally, silver wire (diameter, 120 µm) was placed at the interface region, serving as the shadow mask for Au evaporation. 50-nm Au was evaporated as the electrodes.

Characterizations

Optical imaging. The bright-field optical images were collected by a custom microscope (Olympus BX53).

Photoluminescence imaging and spectra collection. Samples were excited with a light source (012-63000; X-CITE 120 REPL LAMP). The filter cube contains a bandpass filter (330–385 nm) for excitation, and a dichroic mirror (cutoff wavelength, 400 nm) for light splitting and a filter (long pass 420 nm) for emission. The photoluminescence spectra were collected by a spectrometer (SpectraPro HRS-300).

Interdiffusion calculation. We used a simplified one-dimensional diffusion model of Fick's second law to describe the transient concentration profile across the lateral heterostructures. The model is mathematically expressed as $\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D(c)\frac{\partial C}{\partial x}\right)$. Here, C is the concentration of Br, t is the heating time, x is the length and D is the inter-diffusion coefficient of halides. The diffusion coefficient is calculated using the classical Boltzmann–Matano method. The evolution of the concentration profile for the BA and 2T lateral heterostructures is fitted by a normal cumulative distribution function, $y = y_0 + A\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}w}\exp\left(-\frac{(t-x_c)^2}{2w^2}\right)\mathrm{d}t$. The fitted curves, along with the experimental Br concentration obtained from the photoluminescence emission peaks, are shown in Supplementary Fig. 7. We note that the concentration profile obtained using the edge widths and photoluminescence emissions can only be used for a rough estimation of the diffusion coefficients because no gradient is assumed across the overlaid edge widths of the lateral heterostructures. It is estimated that the diffusion coefficient of the $(BA)_2PbI_4$–$(BA)_2PbBr_4$ and $(2T)_2PbI_4$–$(2T)_2PbBr_4$ lateral heterostructures are of the order of approximately 10^{-13} $cm^2 s^{-1}$ to 10^{-12} $cm^2 s^{-1}$ and approximately 10^{-16} $cm^2 s^{-1}$ to 10^{-15} $cm^2 s^{-1}$, respectively.

SEM imaging. The backscattering SEM images were collected by a scanning electron microscope (FEI TeneoVS). The acceleration voltage was 1 kV and the acceleration current was 0.1 nA.

Powder XRD measurements. Powder XRD was measured using a powder X-ray diffractometer (Panalytical Empyrean) with a Cu Kα source.

AFM. AFM images were recorded in tapping mode using an atomic force microscope (Bruker MultiMode 8).

Fluorescence-lifetime imaging microscopy measurements. The samples with PMMA (950 PMMA A4, Microchem) coating were measured with a 50 × 0.55 numerical aperture air objective in a confocal laser scanning microscope (LSM 510 NLO AxioVert200M) equipped with a tunable laser (Mai-Tai HP). The excitation wavelength was 405 nm, from the second harmonic of 810 nm (<100 fs, 80 MHz). Lifetime measurements were performed using time-correlated single photon counting (Becker-Hickl SPC-150). The lifetime decay was collected and analysed using Becker-Hickl SPCM software.

Single-crystal XRD measurement. Single crystals of $(2T)_2PbBr_4$ were analysed using a diffractometer (Bruker Quest) with kappa geometry, an I-µ-S microsource X-ray tube, a laterally graded multilayer Göbel mirror single crystal for monochromatization, an area detector (Photon2 CMOS) and a low-temperature device (Oxford Cryosystems). Examination and data collection were performed with Cu Kα radiation (λ = 1.54178 Å) at 150 K.

Electric measurement. Current–voltage characteristics were acquired by sweeping the bias voltages from −10 V to 10 V and then from 10 V to −10 V based on a probe station (PS100 Lakeshore) with a source meter (Keithley 2400).

TEM characterizations

TEM sample preparation of pure $(2T)_2PbI_4$, $(2T)_2PbBr_4$ sheets and $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructure. To transfer pure $(2T)_2PbI_4$, $(2T)_2PbBr_4$ sheets or $(2T)_2PbI_4$–$(2T)_2PbBr_4$ heterostructures to the TEM grids, a layer of PMMA was spin-coated on the growth SiO_2/Si substrates before perovskite growth. The spin-coating speed was 3,000 rpm and the duration time was 1 min. Then the Si substrates were placed on a 70 °C hot plate for 10 min to condense the PMMA and remove the residue solvent. After that, the perovskites were grown on the PMMA-coated Si.

The transfer process of perovskites from the growth PMMA/Si substrates was performed in the glove box. First, the growth substrates were dipped into CB to dissolve the sacrificial layer PMMA, which facilitates the detachment of samples from the growth substrates. Then, vortex mixing or ultrasonication was used to further detach the samples from Si substrates. Afterwards, the suspension was dropped onto the TEM grids. TEM grids were rinsed with clean CB a few times to remove the residue PMMA. The TEM grids (Lacey/carbon grids on 200-mesh Cu, Ted Pella) were covered by one layer of carbon-nanotube film and mono-layer graphene hybrids to mitigate electron-beam damage.

TEM imaging and spectrum acquisition. The diffraction patterns were obtained on a 200-kV transmission electron microscope (JEOL JEM-2100plus) with a camera (TEMCam-XF416, TVIPS). The AC-HRTEM and EDS mapping was taken on an 80-kV aberration-corrected transmission electron microscope (JEOL GrandARM) equipped with a fast camera (OneView IS, Gatan). The EDS map is an intensity map based on X-ray counts, which are proportional to the content of the element. A low accelerating voltage was used to enhance the contrast at low magnifications when searching for samples and locating the interface positions. The low-dose-rate imaging was achieved by increasing electron-beam spot size (decreasing the beam current) and reducing the brightness of the electron beam (spreading the beam illumination). To avoid unnecessary electron-beam damage, the minimum-dose system was used to reduce damage while searching for the samples and during focusing.

There are usually minor displacements between continuously captured images due to the mechanical vibration of the specimen holder

and the effect of electron irradiation; if we were to simply superpose these images, it would lower the spatial resolution of the HRTEM image. To correct these displacement drifts, we use a digital micrograph script developed by D. R. G. Mitchell (http://www.dmscripting.com/stack_alignment.html). The core of the script is to measure the drift by cross-correlation, thus applying measured drift to each image. Each drift measurement will produce a cross correlation value, which is a guide to whether it is working well ('1' is good, '0.5' is OK). In our case, the cross-correlation values are 0.644, 0.645 and 0.647 when processing the four raw images. After accurate drift alignment, the images can be superposed to produce a HRTEM image with an improved signal-to-noise ratio.

TEM simulation and structural model. The kinematical SAED patterns were simulated with the MacTempasX[44] software. The structural models were constructed with the VESTA 3D visualization program[45].

Molecular dynamics simulations

For the molecular dynamics simulation of the heterojunction between Pb–Br and Pb–I organic–inorganic perovskites, the MYP model for hybrid perovskites was used[46–48]. The MYP model was originally developed for Pb–I perovskites and has been extended to Pb–Br perovskites. It treats the organic–organic interactions by the standard assisted model building with energy refinement (AMBER) force fields, the inorganic–inorganic interactions between Pb, I and Br by Buckingham potentials, and the organic–inorganic interactions as the sum of Buckingham, electrostatic and Lennard–Jones 12-6 terms. As developed and reported, the charges on the metal, halide and cation are non-integer. For our simulations, we employed integer values for these charges; the MYP non-bonded parameters were appropriately rescaled to reproduce the cohesive energy density of the unscaled simulation.

The surface cation geometries were optimized using the program ORCA[49] with the ωB97X-D[50]/def2-TVZP[51,52] density functional theory (DFT) potentials. The standard general AMBER force field (GAFF) parameterization was used for the organic cations[53], and the cation point charges were fit against the electrostatic DFT potential (ωB97X-D/def2-TVZP) of the isolated cation with a +1 charge. The molecular dynamics software LAMMPS was used for the molecular simulations[54]. All simulations used a 1 fs integration timestep and periodic boundary conditions. The particle–particle–particle–mesh (PPPM) algorithm was used for the long-range electrostatics, and Lennard–Jones interactions were truncated at 15 Å. Following the MYP model, the 1-4 pairwise Lennard–Jones interactions were scaled by 0.5 and the 1-4 electrostatics interactions were scaled by 0.8333. A sample LAMMPS input file with all force-field parameters is supplied in Supplementary Data.

The initial perovskite heterojunction geometry was generated by constructing an ideal perovskite monolayer composed of two domains: a $6\times1\times12$ Pb–I domain and a $6\times1\times12$ Pb–Br domain, for a total system size of $12\times1\times12$ of A_2BX_4 unit cells. Halide vacancies were introduced by randomly removing two halide atoms. The simulation was first relaxed under the *NVE* ensemble for 50 ps with restrained atomic displacements of 0.01 Å per timestep. The system was then simulated for 20 ns with the Nosé–Hoover thermostat and barostat, with the barostat applied only to the *x* and *z* directions (which define the plane of the perovskite), to allow the surface cations and halides to be displaced from the perovskite layer, at both 298 K and 800 K. To calculate the free energy required to remove a surface cation and halide, a $6\times1\times6$-unit-cell simulation on pristine (that is, no vacancies) Pb–Br and Pb–I perovskites was run, with a surface cation and a halide atom displaced from their initial positions out to vacuum in 0.5 Å intervals. The system was equilibrated for 0.5 ns and then simulated for 10 ns under the *NVT* ensemble at 298 K for each displacement. The *pymbar* package was used to calculate the free energies[55].

Data availability

All data related to this study are available from the corresponding author on reasonable request.

44. O'Keefe, M. A. & Kilaas, R. *Advances in High-resolution Image Simulation*. Report no. LBL-24727 (Lawrence Berkley National Laboratory, 1988); https://escholarship.org/uc/item/6qb303ch.
45. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Cryst.* **44**, 1272–1276 (2011).
46. Mattoni, A., Filippetti, A., Saba, M. I. & Delugas, P. Methylammonium rotational dynamics in lead halide perovskite by classical molecular dynamics: the role of temperature. *J. Phys. Chem. C* **119**, 17421–17428 (2015).
47. Mattoni, A., Filippetti, A. & Caddeo, C. Modeling hybrid perovskites by molecular dynamics. *J. Phys. Condens. Matter* **29**, 043001 (2017).
48. Hata, T., Giorgi, G., Yamashita, K., Caddeo, C. & Mattoni, A. Development of a classical interatomic potential for $MAPbBr_3$. *J. Phys. Chem. C* **121**, 3724–3733 (2017).
49. Neese, F. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 73–78 (2012).
50. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
51. Schäfer, A., Horn, H. & Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *J. Chem. Phys.* **97**, 2571–2577 (1992).
52. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
53. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general AMBER force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
54. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
55. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105 (2008).

Acknowledgements This work is supported by the Office of Naval Research (grant no. N00014-19-1-2296, programme managers P. Armistead and J. Parker), the National Science Foundation (grant no. 1939986-ECCS, programme manager P. Lane), and at Purdue University, the Davidson School of Chemical Engineering, College of Engineering, and the Birck Nanotechnology Center. TEM work is supported by funding from the National Science Foundation of China (grant no. 21805184), the National Science Foundation Shanghai (grant no. 18ZR1425200) and the Center for High-resolution Electron Microscopy (ChEM) at ShanghaiTech University (grant no. EM02161943). P.Y. acknowledges support from the US Department of Energy, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, under contract no. DE-AC02-05CH11231. J.K acknowledges support from the Air Force Office of Scientific Research (FATE MURI, grant no. FA9550-15-1-0514). B.M.S. acknowledges support from the Air Force Office of Scientific Research (grant no. FA9550-18-S-0003, programme manager K. Caster). We thank L. Huang, B. Boudouris and S. Li for discussions.

Author contributions E.S. synthesized and characterized the 2D perovskite materials; B.Y. and Y.Y. performed TEM characterization and data analysis; S.B.S. and B.M.S. performed molecular dynamics simulations and data analysis; Y. Gao performed organic ligand synthesis; A., Y. Guo, C.S., M.L., P.Y. and J.K. participated in materials characterization and data analysis; E.S. and L.D. wrote the manuscript; all authors read and revised the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2219-7.
Correspondence and requests for materials should be addressed to B.M.S., Y.Y. or L.D.
Peer review information *Nature* thanks Humberto Gutierrez, Hua Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

Article

Late-stage oxidative C(sp^3)–H methylation

https://doi.org/10.1038/s41586-020-2137-8

Received: 6 December 2019

Accepted: 5 March 2020

Published online: 16 March 2020

Check for updates

Kaibo Feng[1,4], Raundi E. Quevedo[1,4], Jeffrey T. Kohrt[2], Martins S. Oderinde[3], Usa Reilly[2] & M. Christina White[1]✉

Frequently referred to as the 'magic methyl effect', the installation of methyl groups—especially adjacent (α) to heteroatoms—has been shown to dramatically increase the potency of biologically active molecules[1–3]. However, existing methylation methods show limited scope and have not been demonstrated in complex settings[1]. Here we report a regioselective and chemoselective oxidative C(sp^3)–H methylation method that is compatible with late-stage functionalization of drug scaffolds and natural products. This combines a highly site-selective and chemoselective C–H hydroxylation with a mild, functional-group-tolerant methylation. Using a small-molecule manganese catalyst, Mn(CF_3PDP), at low loading (at a substrate/catalyst ratio of 200) affords targeted C–H hydroxylation on heterocyclic cores, while preserving electron-neutral and electron-rich aryls. Fluorine- or Lewis-acid-assisted formation of reactive iminium or oxonium intermediates enables the use of a mildly nucleophilic organoaluminium methylating reagent that preserves other electrophilic functionalities on the substrate. We show this late-stage C(sp^3)–H methylation on 41 substrates housing 16 different medicinally important cores that include electron-rich aryls, heterocycles, carbonyls and amines. Eighteen pharmacologically relevant molecules with competing sites—including drugs (for example, tedizolid) and natural products—are methylated site-selectively at the most electron rich, least sterically hindered position. We demonstrate the syntheses of two magic methyl substrates—an inverse agonist for the nuclear receptor RORc and an antagonist of the sphingosine-1-phosphate receptor-1—via late-stage methylation from the drug or its advanced precursor. We also show a remote methylation of the B-ring carbocycle of an abiraterone analogue. The ability to methylate such complex molecules at late stages will reduce synthetic effort and thereby expedite broader exploration of the magic methyl effect in pursuit of new small-molecule therapeutics and chemical probes.

The introduction of methyl groups has the potential to dramatically improve the biological activities of a drug candidate by altering its binding affinity, solubility and metabolism[1–8]. Such changes have been demonstrated to increase the potency of lead compounds more than 2,000-fold and to enable interrogation of biological processes[6–8] (Fig. 1a). Although methyl groups are ubiquitous in small-molecule drugs[1], no general method is available to incorporate them into complex molecules at late stages. Accordingly, de novo synthesis—a rate-limiting step in drug discovery that impairs its overall atom economy—is required[9,10]. A practical synthetic method that allows selective installation of methyl groups from C–H bonds at sites adjacent to heteroatoms, where the magic methyl effect is often most substantial, would streamline the diversification of drug leads and encourage more comprehensive investigations of this effect. Over the past decade, considerable progress has been made in developing C(sp^3)–H alkylation methods in which the N- or O-heterocycle acts as a nucleophilic coupling partner[11–17]. Such cross-couplings have shown broad scope with respect to alkyl electrophiles, but limited scope of the metallated heterocyclic intermediates generated via substrate-controlled deprotonation or single-electron transfer (SET). Cases demonstrated with methyl electrophiles have focused on simple azacycles[11,13,15–17]. Expanding the heterocyclic scope to include dissymmetric substrates, epimerizable stereocentres, electrophilic functionalities (for example, carbonyl and nitrile), remote basic amines, heteroaromatics and halogenated aromatics remains a major challenge to be overcome for widespread use in late-stage diversification. In addition, although direct C–C bond forming reactions may be desirable for installing larger and/or functionalized alkyl groups, direct methylation often results in inseparable mixtures with the starting material owing to the small size and electron neutrality of the methyl group.

We sought to approach C(sp^3)–H methylation in N- and O-containing heterocycles in an oxidative fashion through a hydroxylated intermediate, with subsequent iminium or oxonium ion formation and methylation (Fig. 1b). Catalyst control could be leveraged to influence the site-selectivity and chemoselectivity of C–H hydroxylation in a broad range of heterocycles (Fig. 1c, d). Reports of alkylations of

[1]Department of Chemistry, Roger Adams Laboratory, University of Illinois, Urbana, IL, USA. [2]Pfizer Worldwide Research and Development, Groton Laboratories, Groton, CT, USA. [3]Research and Development, Bristol-Myers Squibb Company, Lawrenceville, NJ, USA. [4]These authors contributed equally: Kaibo Feng, Raundi E. Quevedo. ✉e-mail: mcwhite7@illinois.edu

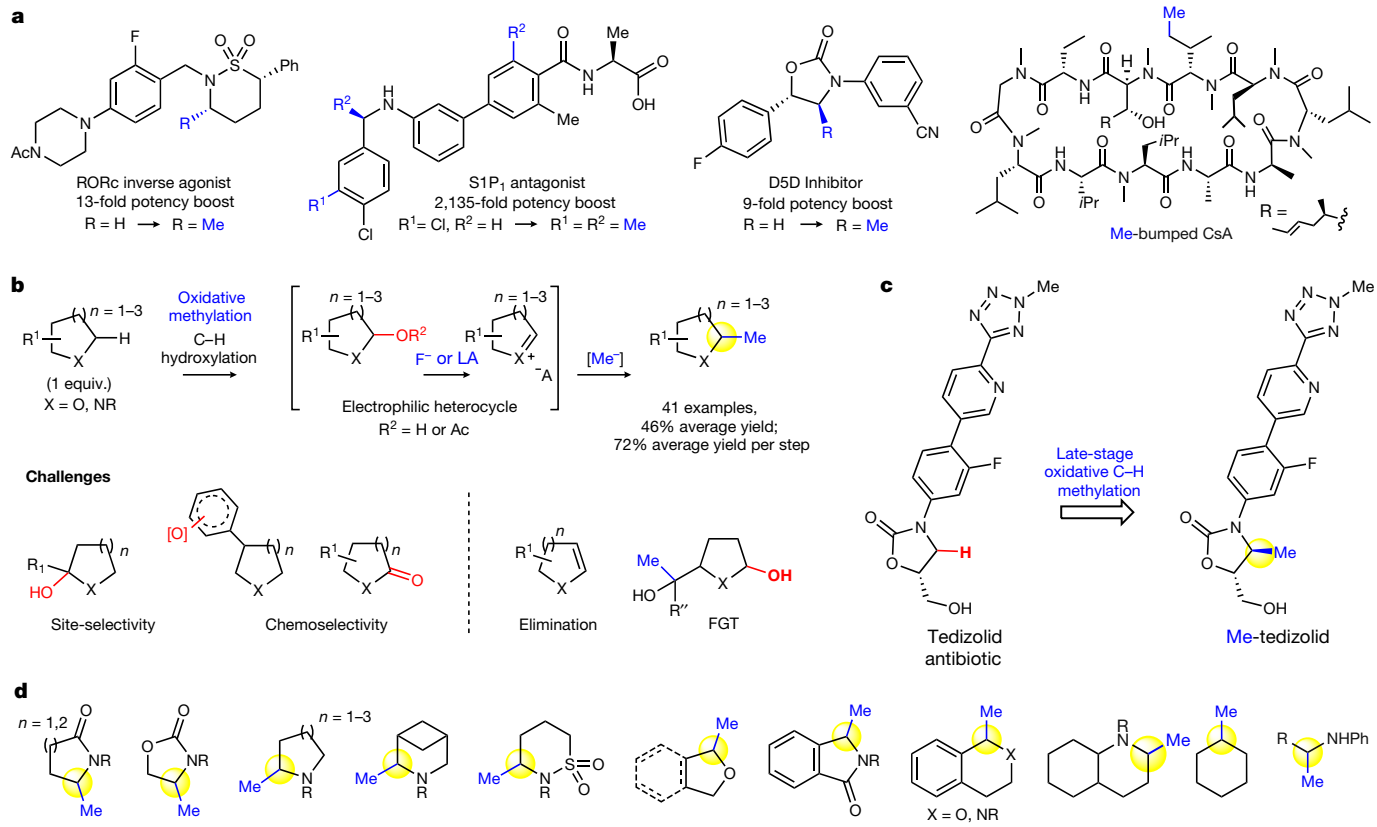
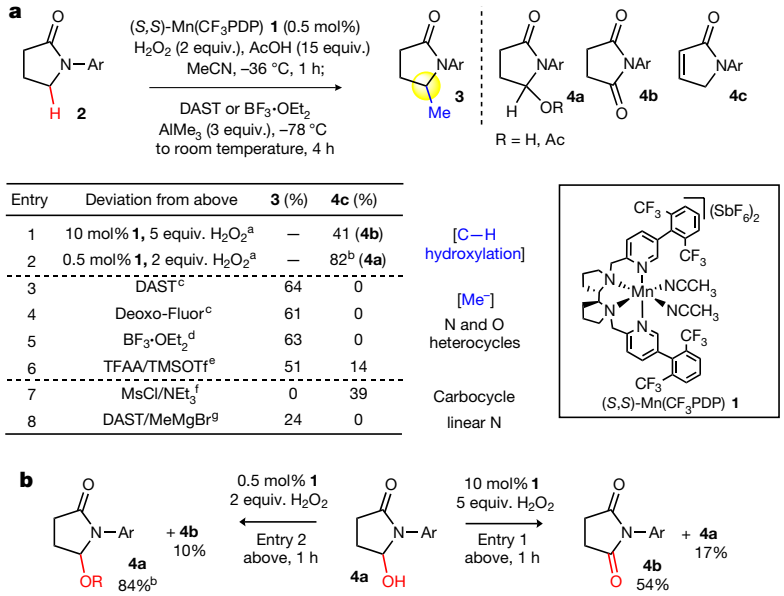


Fig. 1 | C(sp^3)–H methylation. a, The magic methyl effect boosts the potency of drugs (such as RORc, $S1P_1$ and Δ−5 desaturase (D5D)) and furnishes biological probes (such as cyclosporin A (CsA)). **b**, Top row, this oxidative methylation proceeds through an electrophilic intermediate. Bottom row, challenges include unselective oxidation and overoxidation (see site-selectivity and chemoselectivity), as well as elimination and unselective methylation pathways (see functional group tolerance (FGT)). **c**, Late-stage oxidative methylation of the antibiotic tedizolid. **d**, We have demonstrated this oxidative C–H methylation on 16 different pharmaceutically relevant cores. Using one equivalent of substrate, methylation proceeds site-selectively and with functional group tolerance to afford preparative yields in 41 examples (including 18 complex bioactive molecules).

N-acyliminium ions are of limited scope[18,19]. Although C(sp^3)–H oxidation α to heteroatoms has been well demonstrated, substrate-controlled selectivities can afford poor site- and chemoselectivity, thereby limiting examples in complex settings[20]. Moreover, the strong hyperconjugative activation of hemiaminals and hemiacetals typically promotes overoxidation to the corresponding carbonyl, calling for reduction before or after methylation (Fig. 1b)[21–26].

The catalyst $Mn(CF_3PDP)(MeCN)_2(SbF_6)_2$ (**1**) (where CF_3PDP is 1,1′-bis((5-(2,6-bis(trifluoromethyl)phenyl)pyridin-2-yl)methyl)-2,2′-bipyrrolidine) has been reported to uniquely control site- and chemoselectivity in hydroxylating strong methylene C(sp^3)–H bonds while tolerating halogenated arenes, although the tolerance for electron-neutral or electron-rich aromatic and some heteroaromatic rings remained a challenge[27] (Fig. 1b). We questioned whether sterically hindered catalyst **1** could result in faster C–H hydroxylation than alcohol oxidation for hyperconjugatively activated C–H bonds, and whether under milder oxidation conditions such a rate difference would increase chemoselectivity and yield for the hydroxylated product. Under the previously reported forcing conditions (10 mol% **1**, 5 equiv. H_2O_2)[27], oxidation of arylated γ-lactam (**2**) afforded a substantial amount of overoxidation to the corresponding imide (Fig. 2a, **4b**, 41%). Lowering the catalyst and hydrogen peroxide loadings (to 0.5 mol% **1**, 2.0 equiv. H_2O_2) enabled the C–H bond α to nitrogen (α-N) to be hydroxylated to hemiaminal intermediates (hemiaminal, and hemiaminal acetate from AcOH) with an excellent yield of 82% (**4a**). Consistent with slower alcohol oxidation, exposure of alcohol **4a** to identical oxidation conditions afforded 84% hemiaminals with only 10% imide **4b** (Fig. 2b). By contrast, exposure of alcohol **4a** to the forcing conditions afforded predominantly imide **4b** (54%) with only 17% hemiaminals. Under mild oxidation conditions, we also observed enhanced chemoselectivities for electron-rich and electron-neutral aromatics and heteroaromatics, probably because of attenuation of similarly higher-energy overoxidation pathways (vide infra). Notably, this constitutes among the highest substrate/catalyst ratios (S/C = 200) reported so far for a preparative C(sp^3)–H hydroxylation reaction in a complex setting. The ability to separate the hydroxylated intermediate before methylation, while not necessary for alkylation, avoids the formation of an inseparable mixture of the methylated product and starting material, often observed in direct methylation[17]. As expected, Fe(PDP) and $Fe(CF_3PDP)$—used previously for oxidative α-arylation of aliphatic peptides[28–30]—gave a complex mixture of aromatic oxidation products (Extended Data Table 1). Mn(PDP), shown to hydroxylate simple linear amides[31], was not reactive enough to promote preparative hydroxylation of **2**, but can be uniquely effective for some sterically hindered substrates (vide infra).

A further challenge with an oxidative methylation approach was to identify a way to activate the hemiaminal/hemiacetal towards attack by a nucleophilic methyl source without resulting in either undesirable elimination to the enamine, or attack at other electrophilic moieties in complex substrates (Fig. 1b). The modestly nucleophilic and Lewis acidic nature of organoaluminium reagents suggested that they could achieve such selective reactivity. Their high affinity for fluorine, coupled with their tolerance of Lewis acids, afforded a means of generating reactive iminium or oxonium species from transient C–F or from C–OH bond ionization[32,33]. High functional-group tolerance was also expected, given the ability of organoaluminium reagents to methylate oxoniums at late stages in the presence of other electrophilic functionalities[34].



Entry	Deviation from above	3 (%)	4c (%)
1	10 mol% **1**, 5 equiv. H_2O_2[a]	—	41 (**4b**)
2	0.5 mol% **1**, 2 equiv. H_2O_2[a]	—	82[b] (**4a**)
3	DAST[c]	64	0
4	Deoxo-Fluor[c]	61	0
5	$BF_3{\cdot}OEt_2$[d]	63	0
6	TFAA/TMSOTf[e]	51	14
7	MsCl/NEt_3[f]	0	39
8	DAST/MeMgBr[g]	24	0

Fig. 2 | Reaction development. a, Optimization of the oxidative methylation reaction. Note that for achiral substrates, catalysts (*R*,*R*)-**1** and (*S*,*S*)-**1** can be used interchangeably. DAST or $BF_3{\cdot}OEt_2$ was generally used as the hydroxyl activator, with trimethylaluminium as the methyl source. Isolated yields are based on the average of two to three experiments. Ar = *p*-Cl(C_6H_4). **b**, Exposure of hemiaminal (**4a**) to mild C–H hydroxylation (0.5 mol% **1**, 2 equiv. H_2O_2), developed herein, produces little overoxidation to the imide. The previous forcing condition (10 mol% **1**, 5 equiv. H_2O_2) results in imide as the major product. [a]No methylation. [b]Mixture of hemiaminal (64–71%) and hemiaminal acetate (13–18%) from AcOH. [c]1 equiv. [d]2 equiv. [e]TFAA, 1 equiv.; TMSOTf, 1 equiv. [f]MsCl, 1 equiv.; NEt_3, 1 equiv.; $NaHCO_3$ wash; $AlMe_3$, 3 equiv.; –78 °C, 2 h; room temperature, 1 h. [g]MeMgBr, 3 equiv.; –78 °C, 3 h.

After substantial experimentation (shown in abbreviated fashion in Fig. 2a), we arrived at a scalable general procedure using either diethylaminosulfur trifluoride (DAST) or boron trifluoride diethyl etherate ($BF_3{\cdot}OEt_2$) as the hydroxyl activator for iminium formation, with $AlMe_3$ as an inexpensive, commercial methylating reagent. Thermally stable bis(2-methoxyethyl)aminosulfur trifluoride (Deoxo-Fluor) could also be used. In general, the fluorine-assisted methylation strategy should be used in substrates containing Lewis basic or electrophilic functional groups, whereas those lacking such functionality can be methylated via the BF_3 activation strategy (Extended Data Table 1). When unreacted hemiaminal acetate is observed, esterification of the hemiaminal by trifluoroacetic anhydride (TFAA) and subsequent activation of both esters with trimethylsilyl triflate (TMSOTf) to furnish the iminium can be used[34]. The ability to vary the ionization method with $AlMe_3$ is essential to the broad scope of this methylation, providing a facile handle to tune reactivity and/or selectivity for a given substrate in cases in which unreacted hemiaminal intermediates or enamine byproducts are observed (vide infra).

We examined alternative activation modes with $AlMe_3$ and other methylating reagents with DAST (Extended Data Table 1). In hemiaminals, base-mediated formation of an activated C–O bond (that is, mesylation) led predominantly to elimination (Fig. 2a). Grignard reagents, even at cryogenic temperatures, afford diminished yields relative to $AlMe_3$, probably because of poor functional-group tolerance. Although ineffective for the methylation of heterocyclic substrates, these reagents can be used to effect methylation in challenging linear secondary amine and carbocyclic substrates (vide infra; **51**, **53**).

We explored oxidative methylation for its capacity to methylate a collection of 22 compounds comprising 10 different heterocyclic cores commonly found in pharmaceuticals (Fig. 3). We successfully carried out a gram-scale methylation of **2** in 71% yield via DAST activation; an ethyl group could also be installed using commercial triethylaluminium (**5**, 51%). Methylated δ-lactam **6** was isolated in 58% yield; analogous to the γ-lactam, under more forcing oxidation conditions δ-lactam gave predominantly imide (60%). For amide **2**, methylated **3** was observed in preparative yields with both DAST and BF_3 ionization (Fig. 2a). However, in oxazolidinones, housing more labile carbonyls, fluorination with DAST furnished substantially higher yields (**7**, 55% versus 10% with BF_3; Extended Data Table 1; **8**, 63%). Pyrrolidine, the fifth most common nitrogen heterocycle in drugs[35,36], undergoes hydroxylation with no substantial overoxidation to pyrrolidinone, followed by BF_3-promoted methylation to afford monomethylated product **9** in 54% yield. Essential for late-stage derivatization and orthogonal to most radical processes, high site-selectivity for methylation at the less sterically hindered methylene site was observed in substrates bearing more activated tertiary (3°) aliphatic, 3° benzylic, and 3° α-carbonyl C–H bonds to afford products in preparative yields (**10**–**13**). Full stereoretention was measured with chiral substrates leading to **12** and **13**, indicating that the high regioselectivity can be attributed to catalyst control of Mn(CF_3PDP) **1** in the C–H cleavage step. Methyl ester, ketone, acetate and nitrile—not well tolerated with strongly nucleophilic methylation reagents—were maintained using DAST activation/$AlMe_3$ methylation (**12**–**15**). Methylation on a 3-phenylpyrrolidine derivative proceeded regioselectively at the methylene site distal from the phenyl group, furnishing the 5-methylated product **16** in useful yield. Such chemoselectivity for electron-neutral aromatics has not been previously demonstrated: at higher catalyst loadings, **1** afforded poor yields and chemoselectivities[27].

In piperidines—the most common nitrogen heterocycle in small-molecule drugs[35,36]—both DAST and BF_3 activation should be tried: enamine formation competes strongly with methylation and depends highly on both the substrate and the mode of hemiaminal activation (Fig. 3). For example, a pipecolic acid derivative gave 2% of the methylated product with 60% enamine byproduct under DAST activation, whereas the BF_3 activation furnished **17** in 47% overall yield. Alternatively, methylation of piperidinyl-2-methyl acetate using BF_3 did not fully convert the hemiaminal to methylated product, whereas DAST activation afforded **18** in 64% yield. Gamma-substituted piperidines are prevalent structures in drugs, such as in paliperidone and paroxetine. An *N*-nosyl intermediate in the synthesis of paliperidone was selectively methylated using the DAST protocol to give **19** in 37% yield with no protection of the benzisoxazole ring γ to nitrogen. However, methylation

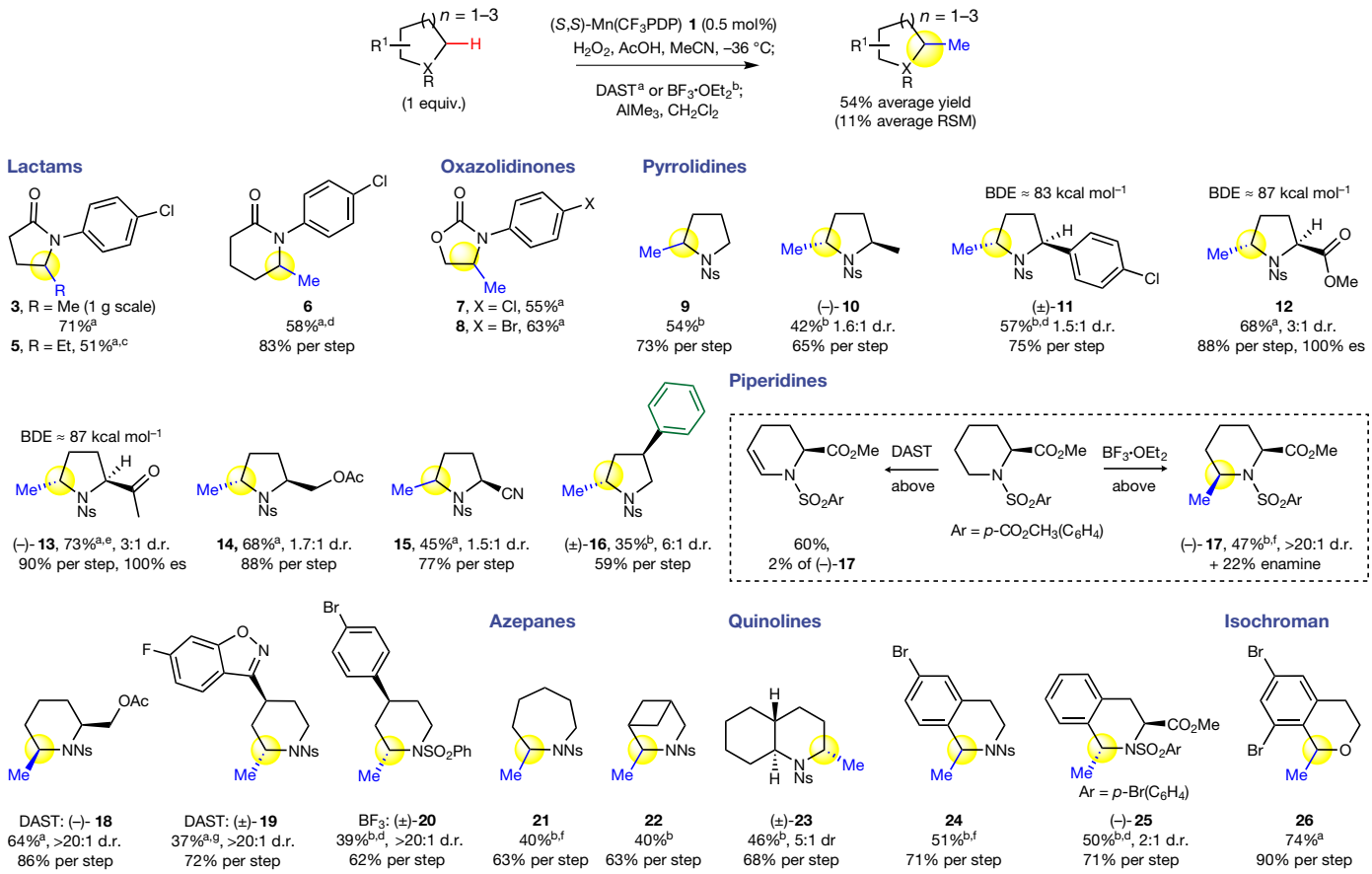


Fig. 3 | Ten different heterocyclic cores, commonly found in pharmaceuticals, explored in the Mn(CF$_3$PDP) (1)-catalysed C–H oxidative methylation. Twenty-two heterocycles—including lactams, oxazolidinones, pyrrolidines, piperidines, azepane, azabicycloheptane, quinolines and isochroman—were oxidatively methylated in preparative overall yields (54% average) using limiting substrate. es, enantiospecificity; RSM, recovered starting material. Isolated yields are based on the average of three experiments. General oxidation: substrate (1 equiv.), catalyst (0.5 mol%), AcOH in MeCN, –36 °C; H_2O_2 (2 or 5 equiv.) in MeCN syringe pump for 1 h. Mixture passed through silica plug, EtOAc flush, concentrated before isolation or methylation. For insoluble substrates, CH_2Cl_2 was added to MeCN and/or the temperature of the oxidation reaction was increased to 0 °C. [a]DAST activation: crude in CH_2Cl_2 (0.2 M), DAST (1 equiv.) added at –78 °C; room temperature (rt) for 1 h; cooled to –78 °C, $AlMe_3$ added, stirred 2 h; rt for 1 h. [b]BF_3 activation: crude in CH_2Cl_2 (0.2 M), –78 °C, $AlMe_3$ (3 equiv.) and BF_3·OEt_2 (2 equiv.) sequentially added, stirred 1 h; rt for 3 h. [c]Triethylaluminium. [d]2 mol% (*S*,*S*)-**1**. [e]$AlMe_3$ –78 °C, 3 h. [f]1 mol% (*S*,*S*)-**1**. [g]For facile purification, hemiaminal isolated before methylation. 10 mol% (*S*,*S*)-**1**, rt, starting material recycled once.

of γ-(4-bromophenyl)piperidine with DAST resulted predominantly in enamine formation, whereas the BF_3 activation strategy afforded 39% yield of methylated **20**. Notably, all piperidines furnished a single observed methylated diastereomer, probably as a result of the rigid half-chair conformation of the iminium intermediate[37].

Other simple cyclic amines—such as azepane, azabicycloheptane and decahydroquinoline—were selectively methylated using the BF_3 activation protocol α-N to afford 40–46% overall yields of mono-methylated products **21**–**23**. Tetrahydroisoquinoline, among the top 20 nitrogen heterocycles in drugs[36], was oxidatively methylated using BF_3 activation in good yields for both a brominated and an unsubstituted aromatic structure (**24**, 51%; **25**, 50%), with lower yields observed using DAST activation. In contrast with most radical-based oxidation methods that oxidize isochromans to isochromanones, using Mn(CF$_3$PDP) **1**, little overoxidation was observed. 6,8-Dibromoisochroman was oxidatively methylated using DAST/$AlMe_3$ in 74% yield (**26**); BF_3 activation for these types of oxygen heterocycles afforded ring-opening products[38].

We explored the ability of highly site- and chemoselective C–H hydroxylation catalysed by Mn(CF$_3$PDP) **1**, coupled to a Lewis-acid/fluorine-promoted methylation, to provide a general method for installing methyl groups directly into the hydrocarbon cores of complex, bioactive molecules, thereby avoiding lengthy and costly de novo synthesis[1,9,10] (Fig. 4). Acetylated cromakalim—a potassium channel activator housing a γ-lactam with tertiary and secondary hyperconjugatively activated α-N C(sp^3)–H bonds—underwent oxidative methylation at the less activated but more sterically accessible secondary site in good yield (**27**, 51%). The acetate could be readily removed to furnish methylated cromakalim **28** in 85% yield. The methyl ester of indoprofen, an anti-inflammatory drug investigated for spinal muscular atrophy[39], was oxidatively methylated at its central isoindolinone core in synthetically useful yields (**29**, 33%). The enhanced chemoselectivity of oxidative methylation with **1** under reduced loadings is evident when comparing with results at higher loadings (10 mol%), in which **29** was obtained in diminished yields (7%) owing to poor chemoselectivity. Chloroindoprofen methyl ester—a derivative with decreased electron density on the aromatic ring—underwent oxidative methylation in higher yields (**30**, 55%).

Fenspiride, an antitussive drug, was oxidatively methylated in a useful overall yield (**31**, 24%) at a methylene site adjacent to the quaternary centre of an unprotected spiro-oxazolidinone, using the (*S*,*S*)-Mn(PDP)$(SbF_6)_2$ catalyst that is less sensitive to sterics[27]. The basic piperidine nitrogen of fenspiride was protected with HBF_4 and rendered a strong electron-withdrawing group, deactivating a distal benzylic site and three α-N sites towards C–H oxidation[40]. Notably, SET reactions proceeding via basic amine catalysis (for example, quinuclidine) are not generally amenable to this kind of nitrogen-protection strategy and

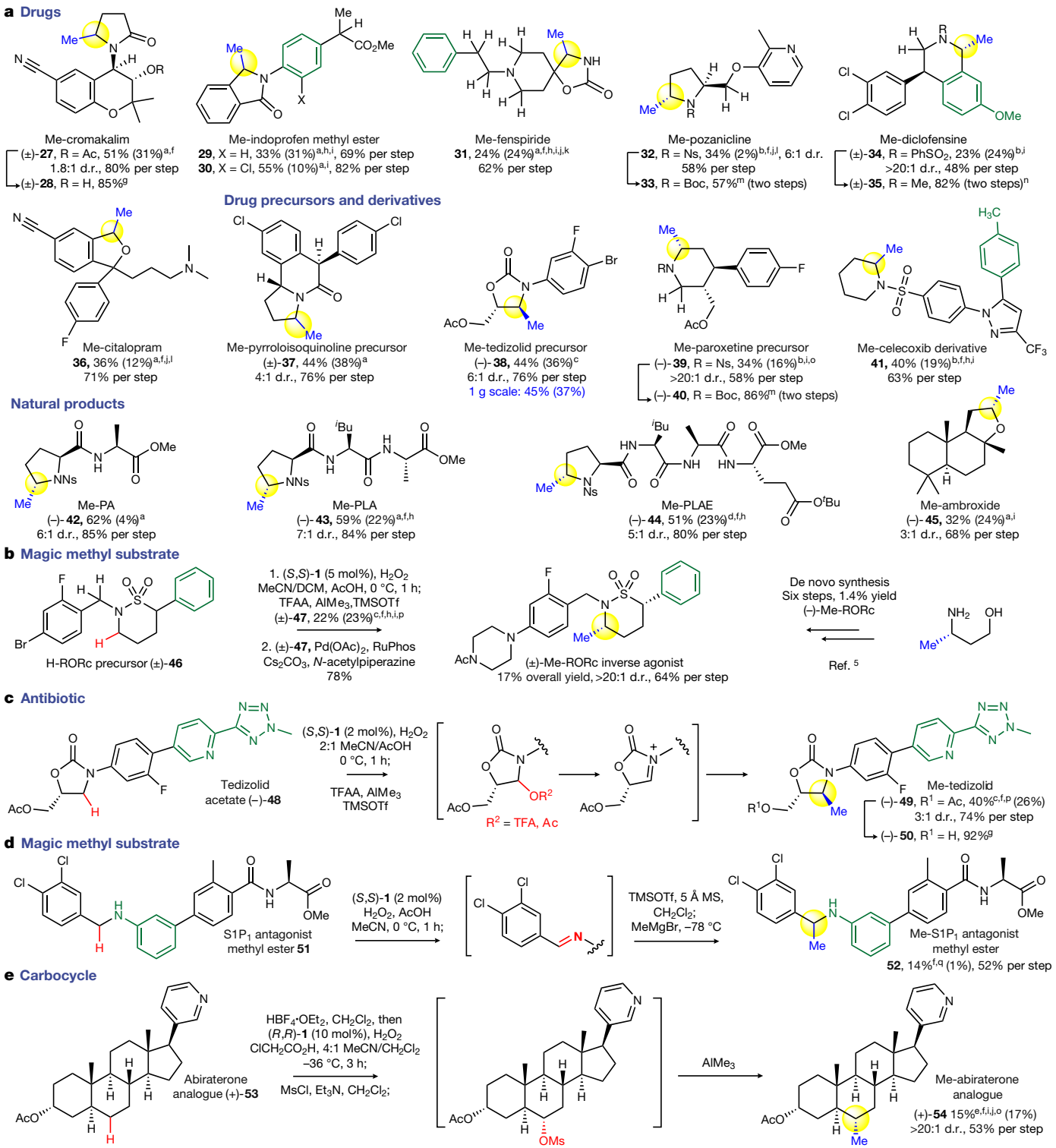


Fig. 4 | Application of oxidative methylation for late-stage functionalization. **a**, Selective methylation of drugs, drug precursors, intermediates and natural products underscores the power of this method for late-stage applications. Generally, 0.5–5 mol% (*S*,*S*)-**1** and 2 or 5 equiv. H_2O_2 were used for oxidation. Higher catalyst and oxidant loadings were applied when conversions were low. Isolated yields are based on the average of three experiments. Explicit C–H bonds denote competing sites of oxidation (for example, Me-fenspiride). Green colouring denotes oxidatively labile aromatic groups (for example, Me-diclofensine). **b**, Methylation of an RORc inverse agonist precursor rapidly furnishes the analogue with a 13-fold boost in potency. Details of synthesis from (*S*)-3-aminobutan-1-ol are from ref. [5]. **c**, Methylation of the antibiotic tedizolid acetate furnishes Me-tedizolid. **d**, Methylation of linear aniline in the $S1P_1$ antagonist methyl ester occurs at a position where the magic methyl effect was observed to contribute to a 2,135-fold potency boost (see ref. [6]). **e**, Remote methylation of a carbocycle on an abiraterone analogue. [a]DAST activation. [b]BF_3 activation. [c]TMSOTf activation: TFAA, rt, 1 h; cooled to −78 °C, $AlMe_3$ and TMSOTf sequentially added, 2 h; then rt, 1 h. [d]Deoxo-Fluor activation. [e]Mesylation activation: MsCl and Et_3N added, rt, 1 h; $NaHCO_3$ wash, dried, condensed; redissolved in CH_2Cl_2, $AlMe_3$ added at −78 °C, stirred 2 h; then rt, 1 h. [f]Oxidation intermediates isolated before methylation. [g]1 M NaOH/MeOH. [h]Starting material recycled once. In some cases, isolated yields are based on an average of two experiments; see Supplementary Information. [i]For insoluble substrates, CH_2Cl_2 was added to MeCN and/or the temperature of the oxidation reaction was increased to 0 °C. [j]HBF_4 protection (see ref. [40]). [k]10 mol% (*S*,*S*)-Mn(PDP)$(SbF_6)_2$. [l]10 mol% (*S*,*S*)-**1**. [m]PhSH, Cs_2CO_3; Boc_2O. [n]Mg, NH_4Cl; formaldehyde, formic acid. [o]10 mol% (*R*,*R*)-**1**. [p]2 equiv. TMSOTf. [q]TMSOTf (1.2 equiv.), 0 °C, 1 h, then MeMgBr (3.0 equiv.) −78 °C, 4 h, repeated once.

therefore do not undergo remote C–H functionalizations[17,24]. A derivative of pozanicline—a neuroprotective drug evaluated for the treatment of attention deficit hyperactivity disorder (ADHD)[41]—undergoes α-N oxidative methylation at the pyrrolidine in useful yield and diastereoselectivity (**32**, 34%, diastereomeric ratio (d.r.) 6:1). The HBF_4 protection deactivates the basic pyridine moiety and its proximal ethereal sites from oxidation with **1**. Although DAST activation produced similar yields, a higher diastereoselectivity was obtained with BF_3 activation (6:1 compared with 3:1), possibly because of different iminium counterions (Fig. 1b). The nosyl (Ns) group on pyrrolidine, a convenient chromophoric protecting group for secondary amines, was readily removed using thiophenol and subsequently protected with tert-butyloxycarbonyl (Boc) to afford **33** in 57% overall yield. Underscoring the unique chemoselectivity of this method, a derivative of the antidepressant diclofensine was oxidatively methylated at its tetrahydroisoquinoline core to afford **34** in useful yield despite the presence of a very electron-rich methoxyphenyl. Mild, reductive desulfonation followed by reductive amination furnished methyl-diclofensine **35** in 82% yield. The antidepressant drug citalopram, upon HBF_4 protection of the tertiary amine, is oxidatively methylated at its dihydroisobenzofuran core to afford **36**. DAST activation was used for most of these densely functionalized substrates, whereas BF_3 activation was more effective for the tetrahydroisoquinoline core.

A precursor to pyrroloisoquinoline—a prevalent structure in compounds with neurotransmitter-uptake-inhibitory properties[42]—undergoes selective oxidative methylation at the less sterically hindered methylene site, versus the more activated tertiary, benzylic sites, to furnish **37** (44% yield, Fig. 4a). Oxidation of a carbamate precursor to the antibiotic tedizolid furnished substantial amounts of hemiaminal acetate that could be methylated in a useful overall yield under TFAA/TMSOTf-assisted methylation (**38**, 44%). This method is operationally facile and can be performed on a gram scale with no loss in efficiency (45% yield). Fluorination afforded lower yields of methylated product **38** owing to unconverted hemiaminal acetate. The core piperidine of a paroxetine precursor and metabolite[43] was oxidatively methylated in useful overall yields (**39**, 34%) preferentially at the less sterically hindered methylene site remote from the 3-acetoxymethyl group. Nosyl deprotection and subsequent Boc protection afforded **40** in 86% yield. A piperidine derivative of the anti-inflammatory drug celecoxib was mono-methylated to afford **41** in good overall yield in the presence of an oxidatively labile tolyl group and pyrazole, both tolerated during C–H oxidation with **1** and requiring no protection. In these piperidine substrates, BF_3 activation was effective in furnishing methylated products.

Methylation of proline-based di-, tri-, and tetrapeptides proceeded with good overall yields and mass balances (**42**, **43**, **44**) with **1** under fluorine-assisted oxidative methylation conditions (Fig. 4a). Deoxo-Fluor may be used in substrates such as that of tetrapeptide **44**, for which isolation from the polar byproducts of DAST is challenging. Although effective in promoting the arylation of peptides with electron-rich aromatic nucleophiles, BF_3 activation in the $AlMe_3$ methylation of peptides afforded complex mixtures, probably arising from activation of the amide carbonyls[30]. Ambroxide, a naturally occurring terpenoid, also underwent selective oxidative methylation using DAST at a methylene site α to oxygen on its tetrahydrofuran ring, affording **45** in 32% yield (and 19% of sclareolide lactone). The use of BF_3 in this case promoted ring opening. Notably, Fe(PDP), ruthenium-mediated oxidation and **1** under forcing conditions all afforded sclareolide lactone as the major product isolated (see Supplementary Information)[10,21,28].

The sultam ring in **46**—an advanced intermediate of an RORc inverse agonist (Fig. 4b)—was oxidatively methylated with **1**, using TFAA/TMSOTf activation with $AlMe_3$, to afford **47** as the *syn*-diastereomer. Other activation modes, such as BF_3, resulted in deleterious elimination pathways. Notably, an oxidatively labile phenyl moiety and a doubly activated benzylic methylene site were tolerated. Previous installation of the *syn*-methyl group afforded a 13-fold increase in potency relative to the unmethylated version in an assay for the interaction between RORc and the steroid receptor coactivator 1 (SRC1); however, it required a six-step de novo synthesis, resulting in a 1.4% overall yield[5]. This analogue and others are now accessible via cross-coupling with methylated intermediate **47**.

Tedizolid, a commercial oxazolidinone antibiotic for acute bacterial skin infections, bears numerous oxidatively sensitive functional groups such as pyridine, tetrazole and *N*-methyl (Fig. 4c). Oxidation by $Mn(CF_3PDP)$ **1** (2 mol%) of tedizolid acetate **48** proceeded in approximately 53% yield of oxidized products (3:1 hemiaminal:acetate) with no protection of the dense nitrogen functionalities. The major challenge was to identify a procedure to install the methyl group from the hemiaminal intermediates. Activation via fluorination furnished primarily eliminated products not observed on the simpler core structure (**38**, Fig. 4a), whereas BF_3 activation resulted in side reactions. However, under TFAA/TMSOTf activation, elimination was suppressed and the methylated product **49** was obtained in a 40% overall yield (74% per step), comparable to that of the simpler precursor **38**. Deprotection of the acetate in 92% yield afforded methyl-tedizolid **50**, an interesting candidate for future biological evaluation given that a ninefold boost in potency has been reported for similar oxazolidinone cores with methylation at the same position[44] (Fig. 1a).

We questioned whether the scope of this reaction could be extended beyond heterocycles, and found that oxidative C–H methylation is not restricted to substrates that can form iminium or oxonium intermediates: promising reactivity has been observed for both imines and remote alcohols generated via C–H hydroxylation with **1**. An antagonist of the sphingosine-1-phosphate receptor-1 ($S1P_1$), the benzylic and aromatic methylations of which afforded a 2,135-fold increase in potency[6], was methylated in its methyl ester form (**51**, Fig. 4d). Although oxidation of the antagonist was successful without the need for protection of the aniline motif, the resulting imine was much less reactive than an iminium and required a stronger nucleophile than $AlMe_3$. In this case, we found that methylmagnesium bromide at cryogenic temperatures—with TMSOTf activation of the imine—produced the methylated product without eroding the amide and ester functional groups (**52**, 14%, 52% per step).

At higher catalyst loadings, $Mn(CF_3PDP)$ **1** is an effective catalyst for methylene C–H bond hydroxylations[27]. Abiraterone analogue **53** was hydroxylated in approximately 32% yield (with 16% ketone) in one step, without recycling the starting material as required in $Fe(CF_3PDP)$ catalysis[40] (Fig. 4e). In carbocyclic substrates, displacement of a C–F bond or ionization with a Lewis acid is difficult; however, mesylates of such aliphatic alcohols are stable and can be activated by $AlMe_3$ to undergo substitution[45]. By replacing fluorination with mesylation, **53** was successfully methylated as a single observed diastereomer (**54**, 15% overall yield, 53% per step), probably through a carbocation intermediate. To the best of our knowledge, this is the first method that enables such remote methylation at unactivated $C(sp^3)$–H bonds. The discovery of this reactivity underscores the importance of developing methylene oxidations that afford predominantly alcohol products.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2137-8.

1. Schönherr, H. & Cernak, T. Profound methyl effects in drug discovery and a call for new C–H methylation reactions. *Angew. Chem. Int. Ed.* **52**, 12256–12267 (2013).
2. Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016); correction **46**, 1760 (2017).

3. Barreiro, E. J., Kümmerle, A. E. & Fraga, C. A. M. The methylation effect in medicinal chemistry. *Chem. Rev*. **111**, 5215–5246 (2011).
4. Leung, C. S., Leung, S. S. F., Tirado-Rives, J. & Jorgensen, W. L. Methyl effects on protein-ligand binding. *J. Med. Chem*. **55**, 4489–4500 (2012).
5. Fauber, B. P. et al. Discovery of 1-{4-[3-fluoro-4-((3S,6R)-3-methyl-1,1-dioxo-6-phenyl-[1,2] thiazinan-2-ylmethyl)-phenyl]-piperazin-1-yl}-ethanone (GNE-3500): a potent, selective, and orally bioavailable retinoic acid receptor-related orphan receptor c (RORc or RORγ) inverse agonist. *J. Med. Chem*. **58**, 5308–5322 (2015).
6. Quancard, J. et al. A potent and selective $S1P_1$ antagonist with efficacy in experimental autoimmune encephalomyelitis. *Chem. Biol*. **19**, 1142–1151 (2012).
7. Belshaw, P. J., Schoepfer, J. G., Liu, K.-Q., Morrison, K. L. & Schreiber, S. L. Rational design of orthogonal receptor-ligand combinations. *Angew. Chem. Int. Ed. Engl*. **34**, 2129–2132 (1995).
8. Shogren-Knaak, M. A., Alaimo, P. J. & Shokat, K. M. Recent advances in chemical approaches to the study of biological systems. *Annu. Rev. Cell Dev. Biol*. **17**, 405–433 (2001).
9. Blakemore, D. C. et al. Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem*. **10**, 383–394 (2018).
10. White, M. C. & Zhao, J. Aliphatic C–H oxidations for late-stage functionalization. *J. Am. Chem. Soc*. **140**, 13988–14009 (2018).
11. Campos, K. R. Direct sp^3 C–H bond activation adjacent to nitrogen in heterocycles. *Chem. Soc. Rev*. **36**, 1069–1084 (2007).
12. Cordier, C. J., Lundgren, R. J. & Fu, G. C. Enantioconvergent cross-couplings of racemic alkylmetal reagents with unactivated secondary alkyl electrophiles: catalytic asymmetric Negishi α-alkylations of N-Boc-pyrrolidine. *J. Am. Chem. Soc*. **135**, 10946–10949 (2013).
13. Beak, P., Basu, A., Gallagher, D. J., Park, Y. S. & Thayumanavan, S. Regioselective, diastereoselective, and enantioselective lithiation-substitution sequences: reaction pathways and synthetic applications. *Acc. Chem. Res*. **29**, 552–560 (1996).
14. Milligan, J. A., Phelan, J. P., Badir, S. O. & Molander, G. A. Alkyl carbon-carbon bond formation by nickel/photoredox cross-coupling. *Angew. Chem. Int. Ed*. **58**, 6152–6163 (2019).
15. Paul, A. & Seidel, D. α-Functionalization of cyclic secondary amines: Lewis acid promoted addition of organometallics to transient imines. *J. Am. Chem. Soc*. **141**, 8778–8782 (2019).
16. Jain, P., Verma, P., Xia, G. & Yu, J.-Q. Enantioselective amine α-functionalization via palladium-catalysed C–H arylation of thioamides. *Nat. Chem*. **9**, 140–144 (2017).
17. Le, C., Liang, Y., Evans, R. W., Li, X. & MacMillan, D. W. C. Selective sp3 C–H alkylation via polarity-match-based cross-coupling. *Nature* **547**, 79–83 (2017).
18. Hiemstra, H. & Speckamp, W. N. in *Comprehensive Organic Synthesis: Selectivity, Strategy & Efficiency in Modern Organic Chemistry* **vol. 2** ch.4.5 (eds Trost, B. M. & Fleming, I.) 1047–1082 (Pergamon Press, 1991).
19. Li, Z., Bohle, D. S. & Li, C.-J. Cu-catalyzed cross-dehydrogenative coupling: a versatile strategy for C–C bond formations via the oxidative activation of sp^3 C–H bonds. *Proc. Natl Acad. Sci. USA* **103**, 8928–8933 (2006).
20. Andrus, M. B. & Lashley, J. C. Copper catalyzed allylic oxidation with peresters. *Tetrahedron* **58**, 845–866 (2002).
21. Kato, N., Hamaguchi, Y., Umezawa, N. & Higuchi, T. Efficient oxidation of ethers with pyridine N-oxide catalyzed by ruthenium porphyrins. *J. Porphyr. Phthalocyanines* **19**, 411–416 (2015).
22. Ito, R., Umezawa, N. & Higuchi, T. Unique oxidation reaction of amides with pyridine-N-oxide catalyzed by ruthenium porphyrin: direct oxidative conversion of N-acyl-L-proline to N-acyl-L-glutamate. *J. Am. Chem. Soc*. **127**, 834–835 (2005).
23. Yoshifuji, S., Tanaka, K.-I., Kawai, T. & Nitta, Y. Chemical conversion of cyclic α-amino acids to α-aminodicarboxylic acids by improved ruthenium tetroxide oxidation. *Chem. Pharm. Bull*. **33**, 5515–5521 (1985).
24. Kawamata, Y. et al. Scalable, electrochemical oxidation of unactivated C–H bonds. *J. Am. Chem. Soc*. **139**, 7448–7451 (2017).
25. Annese, C., D'Accolti, L., Fusco, C., Licini, G. & Zonta, C. Heterolytic (2e) vs homolytic (1e) oxidation reactivity: N–H versus C–H switch in the oxidation of lactams by dioxirans. *Chem. Eur. J*. **23**, 259–262 (2017).
26. Cui, L., Peng, Y. & Zhang, L. A two-step, formal [4+2] approach toward piperidin-4-ones via Au catalysis. *J. Am. Chem. Soc*. **131**, 8394–8395 (2009).
27. Zhao, J., Nanjo, T., de Lucca, E. C. & White, M. C. Chemoselective methylene oxidation in aromatic molecules. *Nat. Chem*. **11**, 213–221 (2019).
28. Chen, M. S. & White, M. C. Combined effects on selectivity in Fe-catalyzed methylene oxidation. *Science* **327**, 566–571 (2010).
29. Gormisky, P. E. & White, M. C. Catalyst-controlled aliphatic C–H oxidations with a predictive model for site-selectivity. *J. Am. Chem. Soc*. **135**, 14052–14055 (2013).
30. Osberger, T. J., Rogness, D. C., Kohrt, J. T., Stepan, A. F. & White, M. C. Oxidative diversification of amino acids and peptides by small-molecule iron catalysis. *Nature* **537**, 214–219 (2016).
31. Milan, M., Carboni, G., Salamone, M., Costas, M. & Bietti, M. Tuning selectivity in aliphatic C–H bond oxidation of N-alkylamides and phthalimides catalyzed by manganese complexes. *ACS Catal*. **7**, 5903–5911 (2017).
32. Nicolaou, K. C., Dolle, R. E., Chucholowski, A. & Randall, J. L. Reactions of glycosyl fluorides. Synthesis of C-glycosides. *J. Chem. Soc. Chem. Commun*. 1153–1154 (1984).
33. Posner, G. H. & Haines, S. R. Conversion of glycosyl fluorides into c-glycosides using organoaluminum reagents. Stereospecific alkylation at C-6 of a pyranose sugar. *Tetrahedron Lett*. **26**, 1823–1826 (1985).
34. Mason, J. D. & Weinreb, S. M. Total syntheses of the monoterpenoid indole alkaloids (±)-alstoscholarisine B and C. *Angew. Chem. Int. Ed*. **56**, 16674–16676 (2017).
35. Taylor, R. D., MacCoss, M. & Lawson, A. D. G. Rings in drugs. *J. Med. Chem*. **57**, 5845–5859 (2014).
36. Vitaku, E., Smith, D. T. & Njardarson, J. T. Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among U.S. FDA approved pharmaceuticals. *J. Med. Chem*. **57**, 10257–10274 (2014).
37. Stevens, R. V. Nucleophilic additions to tetrahydropyridinium salts. Applications to alkaloid syntheses. *Acc. Chem. Res*. **17**, 289–296 (1984).
38. Tomooka, K., Matsuzawa, K., Suzuki, K. & Tsuchihashi, G. I. Lactols in stereoselection 2. Stereoselective synthesis of disubstituted cyclic ethers. *Tetrahedron Lett*. **28**, 6339–6342 (1987).
39. Lunn, M. R. et al. Indoprofen upregulates the survival motor neuron protein through a cyclooxygenase-independent mechanism. *Chem. Biol*. **11**, 1489–1493 (2004).
40. Howell, J. M., Feng, K., Clark, J. R., Trzepkowski, L. J. & White, M. C. Remote oxidation of aliphatic C–H bonds in nitrogen-containing molecules. *J. Am. Chem. Soc*. **137**, 14590–14593 (2015).
41. Prendergast, M. A. et al. Central nicotinic receptor agonists ABT-418, ABT-089, and (–)-nicotine reduce distractibility in adult monkeys. *Psychopharmacology* **136**, 50–58 (1998).
42. Maryanoff, B. E. et al. Pyrroloisoquinoline antidepressants. Potent, enantioselective inhibition of tetrabenazine-induced ptosis and neuronal uptake of norepinephrine, dopamine, and serotonin. *J. Med. Chem*. **27**, 943–946 (1984).
43. Sugi, K. et al. Improved synthesis of paroxetine hydrochloride propan-2-ol solvate through one of metabolites in humans, and characterization of the solvate crystals. *Chem. Pharm. Bull*. **48**, 529–536 (2000).
44. Fujimoto, J. et al. Discovery of 3,5-diphenyl-4-methyl-1,3-oxazolidin-2-ones as novel, potent, and orally available Δ-5 desaturase (D5D) inhibitors. *J. Med. Chem*. **60**, 8963–8981 (2017).
45. Kitamura, M., Ohmori, K. & Suzuki, K. Divergent behavior of cobalt-complexed enyne having a leaving group. *Tetrahedron Lett*. **40**, 4563–4566 (1999).

Data availability

The data that support the findings of this study are available in the Supplementary Information and from the corresponding author upon reasonable request.

Acknowledgements Financial support for this work was provided by the National Institute of General Medical Sciences (NIGMS) Maximizing Investigators' Research Award (MIRA; grant R35 GM122525), and from Pfizer to study the modifications of natural products and medicinal compounds. We thank L. Zhu and the University of Illinois School of Chemical Science (SCS) nuclear magnetic resonance (NMR) laboratory for assistance with NMR spectroscopy, and B. Budaitis for checking the procedure in Fig. 3, molecule **8**. The Bruker 500-Mz NMR spectrometer was obtained with the financial support of the Roy J. Carver Charitable Trust, Muscatine, IA, USA.

Author contributions K.F. and R.E.Q. conducted the experiments and analysed the data. M.C.W., K.F. and R.E.Q. wrote the manuscript. M.C.W., K.F., R.E.Q., J.T.K., M.S.O. and U.R. designed the project. All authors provided comments on the experiments and manuscript during its preparation.

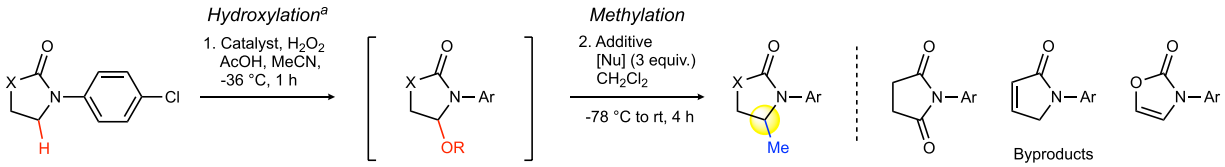
Competing interests The University of Illinois has filed a patent application (number 16/569,492) on the Mn(CF_3PDP) catalyst that lists M.C.W. as an inventor.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2137-8.

Correspondence and requests for materials should be addressed to M.C.W.

Extended Data Table 1 | Reaction optimization



2, X = CH_2
S1, X = O
(1 equiv.)

4a (OH), R = H; **4a (OAc)**, R = Ac, X = CH_2
S2, R = H; **S3**, R = Ac, X = O

3, X = CH_2
7, X = O

4b **4c** **S4**

Entry	Substrate	Catalyst	Loading (mol %)	Additive	[Nu]	**4a (OH)**/ **S2** (%)	**4a (OAc)**/ **S3** (%)	**3**/**7** (%)	**4b** (%)	**4c**/**S4** (%)	rsm (%)
Oxidation											
1[b]	**2**	Fe(PDP)	3 x 5	—	—	< 5[k]	0	—	< 5[k]	—	0
2[c]	**2**	Fe(CF_3PDP)	3 x 5	—	—	8[k]	0	—	6[k]	—	0
3[d]	**2**	Mn(PDP)($OTf)_2$	1	—	—	12	0	—	0	—	75
4	**2**	Mn(PDP)($SbF_6)_2$	1	—	—	28	7	—	< 5[k]	—	35
5[e]	**2**	Mn(CF_3PDP) **1**	10	—	—	13[k]	10	—	41	—	0
6	**2**	**1**	1	—	—	51	21	—	9	—	0
7	**2**	**1**	0.5	—	—	64	18	—	< 5[k]	—	4
Methylation											
8[f]	**2**	**1**	0.5	BF_3•OEt_2	$AlMe_3$	< 5[k]	0	63	< 5[k]	0	11
9[f]	**S1**	**1**	0.5	BF_3•OEt_2	$AlMe_3$	11	5	10	—	4	27
10[g]	**S1**	**1**	0.5	DAST	$AlMe_3$	0	14[k]	55	—	0	16
11[g]	**2**	**1**	0.5	DAST	$AlMe_3$	0	0	64	< 5[k]	0	12
12[g]	**2**	**1**	0.5	Deoxo-Fluor	$AlMe_3$	0	0	61	6	0	5
13[h]	**2**	**1**	0.5	TFAA/ TMSOTf	$AlMe_3$	0	0	51	< 5[k]	14	9
14[h]	**S1**	**1**	0.5	TFAA/ TMSOTf	$AlMe_3$	0	0	46	—	20	13
15[i]	**2**	**1**	0.5	MsCl/NEt_3	$AlMe_3$	15	0	0	< 5[k]	39	6
16[g]	**2**	**1**	0.5	DAST	$ZnMe_2$	17	9	0	11	0	14
17[g,j]	**2**	**1**	0.5	DAST	MeMgBr	24	< 5[k]	24	< 5[k]	0	9

[a]General oxidation (unless otherwise noted): **2** (0.3 mmol), catalyst (x mol%, (*R*,*R*) and (*S*,*S*) enantiomers used interchangeably), AcOH (15 equiv.), MeCN (0.5 M), −36 °C; H_2O_2 (2 equiv.) in MeCN (3.75 ml) syringe pump 1 h. Mixture passed through silica plug, EtOAc flush, concentrated before isolation or methylation. Isolated yields are based on the average of three experiments, unless otherwise noted. [b]Procedure from ref. [28]. [c]Procedure from ref. [29]. [d]Procedure from ref. [31]. [e]5 equiv. H_2O_2. [f]General BF_3 alkylation: crude in CH_2Cl_2 (0.2 M), −78 °C, $AlMe_3$ (3 equiv.) and BF_3•OEt_2 (2 equiv.) sequentially added, stirred 1 h; room temperature (rt) for 3 h. [g]General fluorine alkylation: crude in CH_2Cl_2 (0.2 M), fluorine additive (1 equiv.) added at −78 °C; rt for 1 h; cooled to −78 °C, nucleophile (3 equiv.) added, stirred 2 h; rt for 1 h. [h]General TMSOTf alkylation: crude in CH_2Cl_2 (0.2 M), TFAA (1 equiv.) added, stirred 1 h; cooled to −78 °C, $AlMe_3$ (3 equiv.) and TMSOTf (1 equiv.) sequentially added, stirred 2 h; rt for 1 h. [i]Crude in CH_2Cl_2 (0.2 M), MsCl (1 equiv.) and Et_3N (1 equiv.) added, stirred 1 h; washed $NaHCO_3$, dried, reduced; redissolved in CH_2Cl_2, $AlMe_3$ (3 equiv.) added at −78 °C, stirred 2 h; rt for 1 h. [j]MeMgBr (3 equiv.) added at −78 °C, stirred 3 h. [k]Yield determined by crude ^{1}H-NMR.

Article

Months-long thousand-kilometre-scale wobbling before great subduction earthquakes

https://doi.org/10.1038/s41586-020-2212-1

Received: 17 May 2019

Accepted: 5 February 2020

Published online: 29 April 2020

Check for updates

Jonathan R. Bedford[1]✉, Marcos Moreno[2], Zhiguo Deng[1], Onno Oncken[1,3], Bernd Schurr[1], Timm John[3], Juan Carlos Báez[4] & Michael Bevis[5]

Megathrust earthquakes are responsible for some of the most devastating natural disasters[1]. To better understand the physical mechanisms of earthquake generation, subduction zones worldwide are continuously monitored with geophysical instrumentation. One key strategy is to install stations that record signals from Global Navigation Satellite Systems[2,3] (GNSS), enabling us to track the non-steady surface motion of the subducting and overriding plates before, during and after the largest events[4–6]. Here we use a recently developed trajectory modelling approach[7] that is designed to isolate secular tectonic motions from the daily GNSS time series to show that the 2010 Maule, Chile (moment magnitude 8.8) and 2011 Tohoku-oki, Japan (moment magnitude 9.0) earthquakes were preceded by reversals of 4–8 millimetres in surface displacement that lasted several months and spanned thousands of kilometres. Modelling of the surface displacement reversal that occurred before the Tohoku-oki earthquake suggests an initial slow slip followed by a sudden pulldown of the Philippine Sea slab so rapid that it caused a viscoelastic rebound across the whole of Japan. Therefore, to understand better when large earthquakes are imminent, we must consider not only the evolution of plate interface frictional processes but also the dynamic boundary conditions from deeper subduction processes, such as sudden densification of metastable slab.

Much of the research into earthquake hazard focuses on approximating the magnitudes of the next large earthquakes and estimating when they may occur. Accordingly, regions of the subduction interface that are estimated to have accumulated enough elastic potential energy to match previous earthquakes in that location are considered to be mature segments where a major earthquake can be considered overdue[8]. This hazard assessment is limited by our ability to estimate the long-term relative velocities of the subducting and overriding plates, especially if most of the accumulation period has occurred over times when no modern geophysical instrumentation was deployed. Recently, it has been discovered that relative plate velocities recorded by GNSS vary not only as a postseismic reaction to large earthquakes at neighbouring segments of the subduction plate interface[4–6], but also at mature segments years to days before the onset of main earthquake ruptures[9–11]. Such precursory motions, recorded on the overriding plate surface, are thought to be related to a gradual evolution of stress conditions on the plate interface leading to mainshock failure, which sometimes occurs as a cascading series of foreshocks and respective foreshock afterslips[9] and other times as accelerating aseismic creep[12]. Earthquake precursory phenomena have been extensively studied in laboratory shearing experiments and consist of noticeable changes in acoustic emissions and fluctuation in the measured stress[13,14].

To ascertain the prevalence of tectonic transients in fault zones, we need methods that distinguish the purely tectonic motion from the other non-tectonic signals and artefacts in the continuous GNSS time series[15]. A recently developed time series regression approach called Greedy Automatic Signal Decomposition (GrAtSiD)[7] allows for the automatic estimation of variable interseismic motions at GNSS stations. This technique models the interseismic motions as the sum of a slope, an offset and a sparse number of transient functions (for example, exponential decays; see Methods).

We applied GrAtSiD to continuous GNSS time series in the broad regions of the M_w 8.8 Maule, Chile, 2010 and M_w 9.0 Tohoku-oki, Japan, 2011 earthquakes (hereafter referred to as the Maule and Tohoku-oki earthquakes, respectively) (Fig. 1 and Extended Data Fig. 1), outputting a modelled trajectory for each directional component of the time series that has been separated from jumps (both tectonic and non-tectonic) and steady-state seasonal oscillations (Methods). To interpret this variable interseismic trend it is crucial to identify the sources of non-tectonic signals that may be present in the original time series. By comparing the transient signals of interest (that have been

[1]Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Potsdam, Germany. [2]Departamento de Geofísica, Universidad de Concepción, Concepción, Chile. [3]Institute of Geological Sciences, Freie Universität Berlin, Berlin, Germany. [4]University of Chile, National Seismological Centre, Santiago, Chile. [5]School of Earth Sciences, Ohio State University, Columbus, OH, USA. ✉e-mail: jbed@gfz-potsdam.de

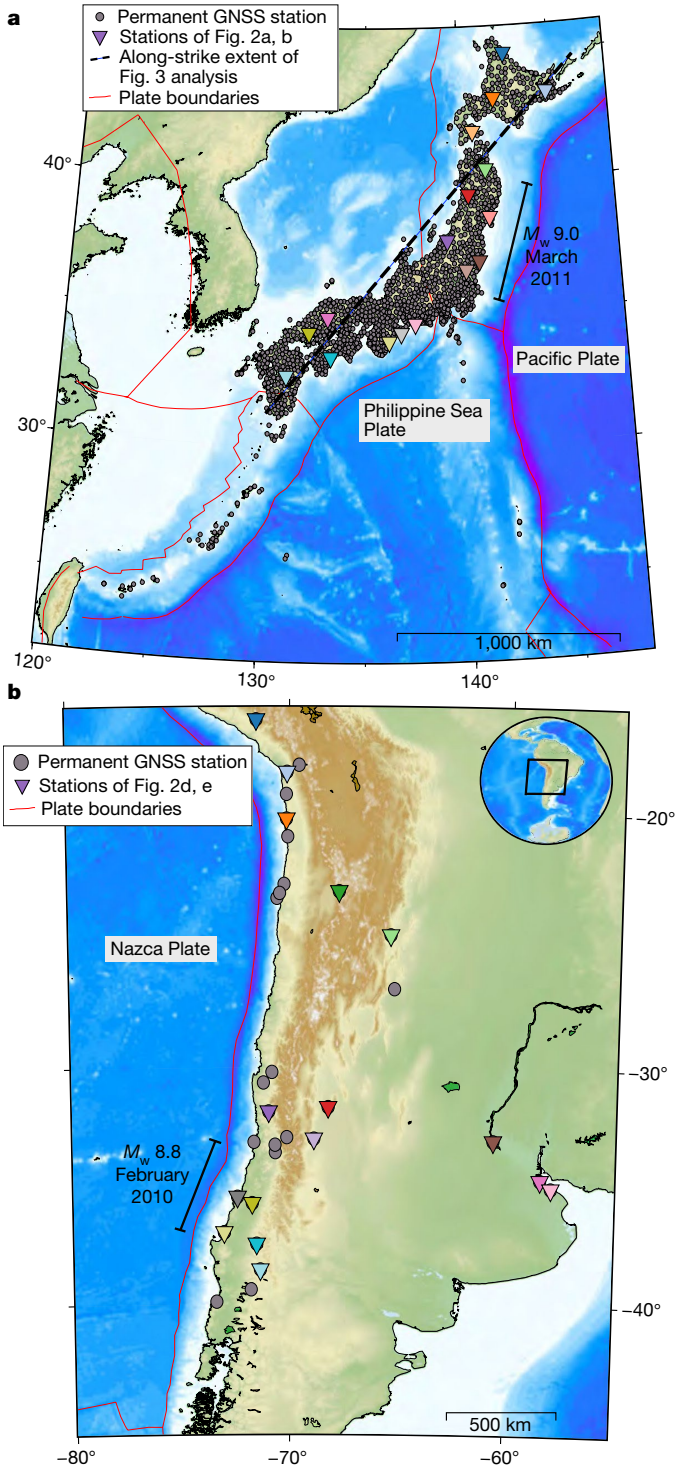


Fig. 1 | Locations of the great earthquakes and of the continuous GNSS stations that capture the preceding transient motions. a, Continuous GNSS stations (grey dots) used in this study. Inverted triangles indicate stations whose time series are shown in Fig. 2 (corresponding to coloured time series). The dashed black line is used in the along-strike analysis of Fig. 3. **b**, As for **a**, but for the South American stations in the region of the Maule earthquake.

isolated by GrAtSiD) to time series predicted from the fluid loading (atmospheric, hydrological and oceanic)[16], and also by analysing the lateral extent of the transient signals (to probe a possible reference frame explanation), we determine that the source of these signals is most likely to be of tectonic origin (Methods).

Features of the modelled trajectories

Figure 2 shows the detrended displacement time series before and after the removal of background seasonal oscillation[17,18] and common-mode errors[19]. We can clearly see that there are variable interseismic rates and strong accelerations in the months leading up to both great earthquakes. The application of GrAtSiD has reduced the noise (seasonal and common-mode) that otherwise obscures this non-steady signal (see also Methods, Extended Data Figs. 2 and 3, and Supplementary Videos 1 and 2). Various transient motions exist in the time series, but the most prominent and long-lived episodes are those in the final 5 and 7 months before the Tohoku-oki and Maule earthquakes, respectively, during which the peak-to-peak amplitude of the horizontal reversals range between 4 mm and 8 mm. These anomalous reversal signals before the Tohoku-oki earthquake are prominent even in the non-filtered time series, whereas before the Maule earthquake, one can clearly see the reversals only after the removal of the seasonal. Before the Tohoku-oki earthquake, there is an earlier, milder reversing signal beginning approximately 19–20 months before the mainshock that shares many spatial characteristics with the later, faster reversals (Supplementary Video 3). Owing to their appearance in the time series, we also refer to these fast reversals in surface motion as 'wobbles'. Before both earthquakes, the approximate azimuth of the horizontal reversal is perpendicular to the strike of the earthquake focal mechanism, which strongly suggests a tectonic origin for the transient signal.

Given the exceptionally dense station spacing in Japan, we were able to perform an along-strike analysis of the transient velocity field along the whole subduction zone (Fig. 3). Although there is some variation of the interseismic velocity along the subduction margin, there is a strong wobbling beginning in October 2010 and lasting until the onset of the Tohoku-oki earthquake in March 2011. A zoom-in on this time window shows that, in the East component, there is a migration of the velocity front across the Earth's surface of approximately 1,000 km in a fortnight (a rate of about 3 km h^{-1}). This velocity front propagates along Japan from the southwest and seems to start as far away as near Taiwan (Supplementary Videos 3 and 4). Because there are fewer stations in the South American network, it is harder to track a velocity front before Maule, although we are able to see a mostly East–West wobbling along Chile in the months preceding the Maule earthquake (Fig. 4, Supplementary Videos 5 and 6). We note also that the Maule reversal in Fig. 4 looks noisier than the Japanese reversal in Fig. 3c. This is because the pre-Maule case has far fewer stations and so we are not able to spatially average the detrended displacements. Before the Tohoku-oki earthquake, anomalous wobbling occurs in both the vertical and horizontal components (Fig. 3, Extended Data Figs. 2 and 4), whereas before the Maule earthquake, the wobbling is mainly in the East–West direction. In the Argentinian stations, there is a strong transient subsidence beginning 4–5 months before the Maule earthquake that persists until the rupture (Figs. 2, 4, Extended Data Fig. 3 and Supplementary Video 6).

Comparison with previous studies

We hereafter refer to the wobbling periods 5 and 7 months before the Tohoku-oki and Maule earthquakes, respectively, as the unstable periods. In previous studies of both mainshocks, precursory seismic periods have been reported: before the Maule earthquake, anomalous shallow (≤40 km) and deep (≥80 km) seismicity occurred within 0.5° of the mainshock hypocentre, beginning approximately 40 days before the mainshock[20]. Before the Tohoku-oki earthquake, a foreshock phase is reported to have begun in late November 2010[21] within 2° of the mainshock epicentre and with an enhanced rate of deeper (≥80 km) seismic moment release at approximately 60 days before the mainshock[20]. Slow slip, inferred from repeating earthquake analysis[21], an onshore volumetric strainmeter[22] and ocean-bottom pressure time series[22], is thought to have begun around 1 month before the Tohoku-oki earthquake and

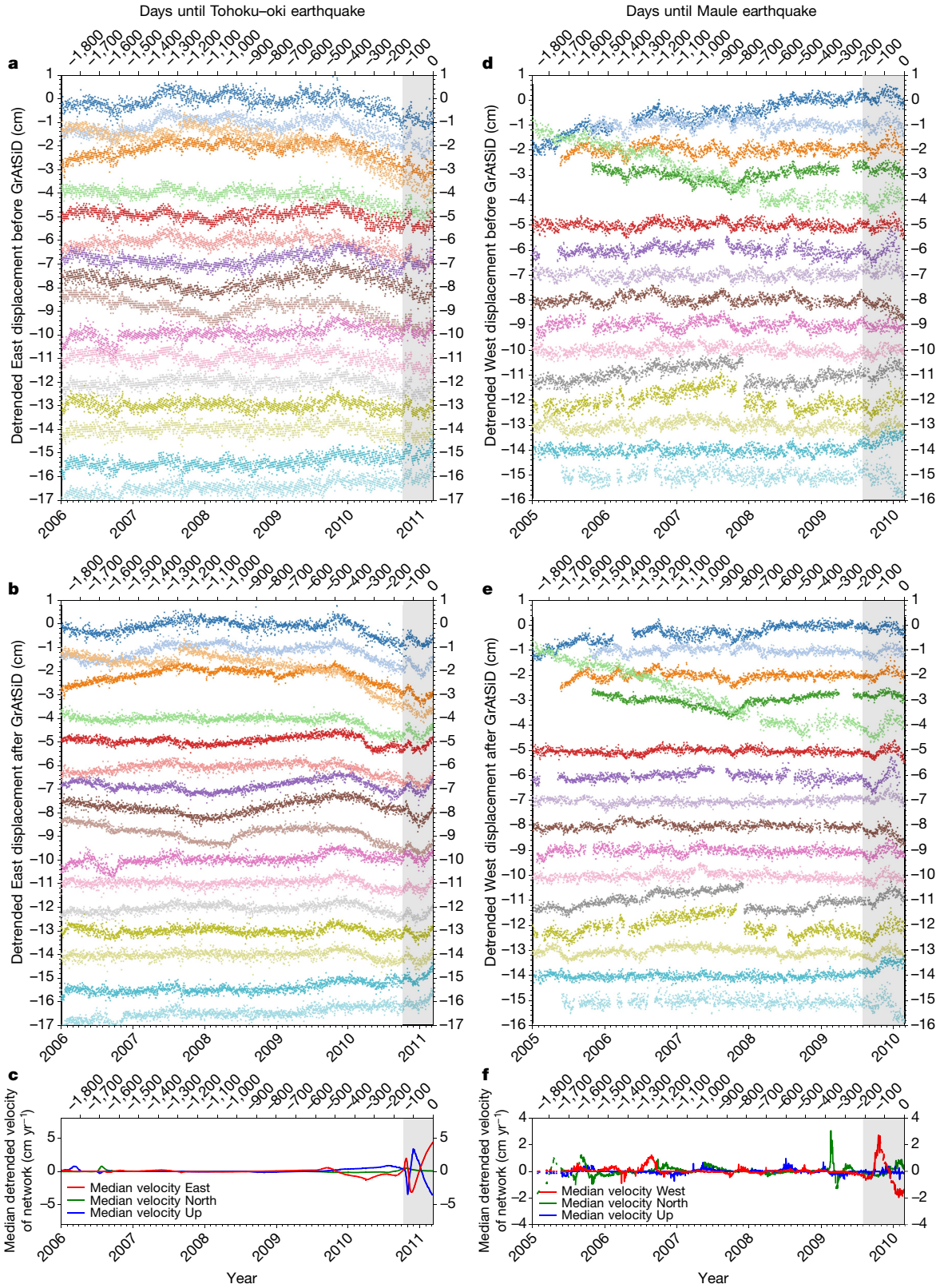


Fig. 2 | Time series before Tohoku-oki and Maule and the effect of noise removal with GrAtSiD. a, East displacement time series before the 2011 Tohoku-oki earthquake (colours going from top to bottom correspond to locations of coloured triangles on Fig. 1 going from north to south). Time series have been manually detrended for optimal display. Offsets (steps) in these time series (automatically solved by GrAtSiD) have been removed. **b**, Time series from panel **a** after removal of background, fluid-loading-induced seasonal oscillations (annual and semi-annual) and network-correlated daily noise (common-mode error). See Extended Data Figs. 2 and 3 for time series before GrAtSiD processing. The grey shading on the right indicates the unstable period before the mainshock. **c**, Average (median) deviation from median velocity at each station of the Japanese network, where median velocity of each station is determined between 1 January 2006 and 8 March 2011. **d**, **e**, As for **a** and **b**, except for the South American stations shown on Fig. 1 and in the West (mainly trenchward) direction. **f**, As for **c**, but for the South American data. Median station velocities calculated between 1 January 2005 and 25 February 2010. We note that in both **c** and **f**, the average deviation from background velocity is only calculated if more than 55% of stations have data on that particular day.

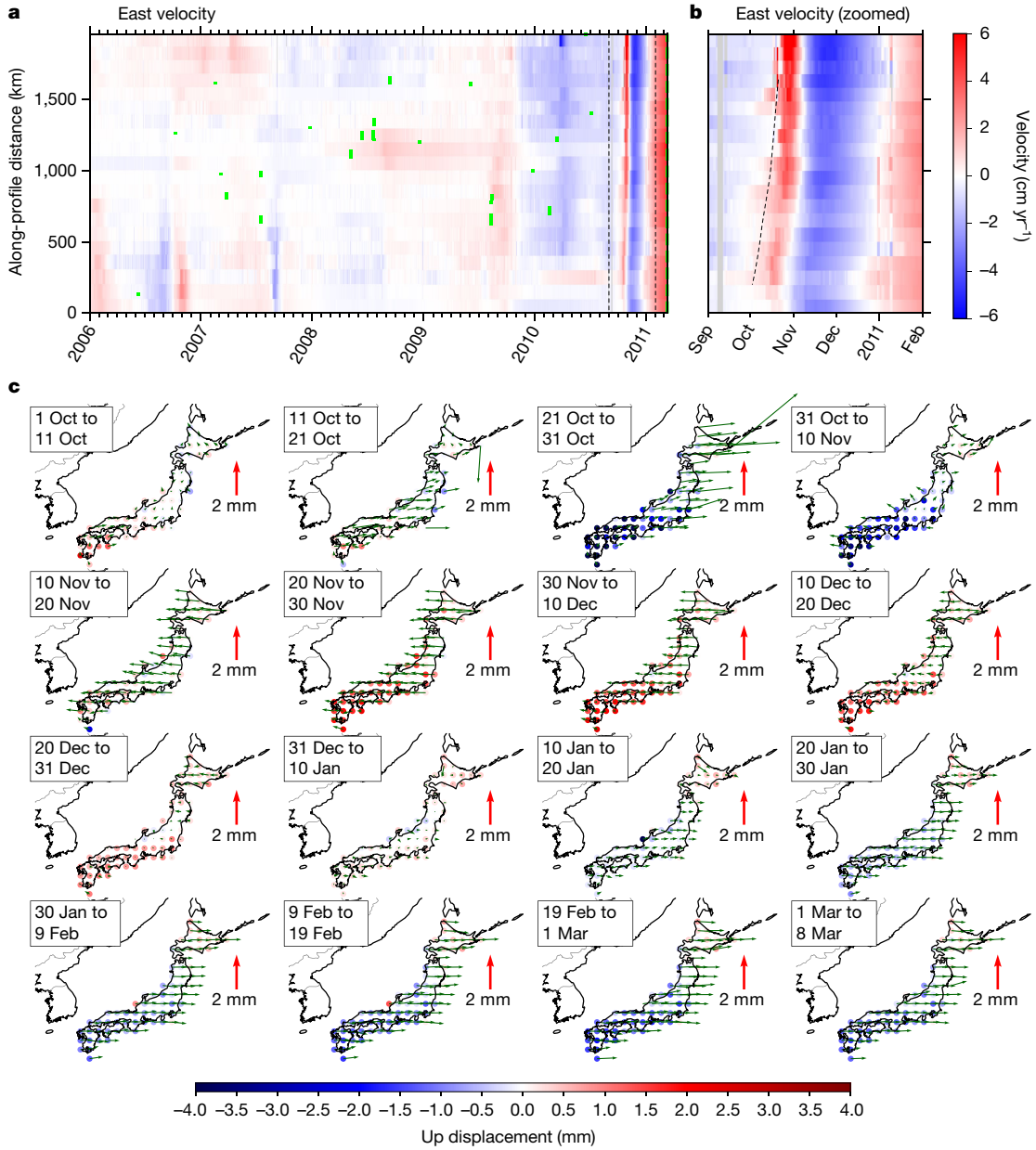


Fig. 3 | Visualizing the along-strike signal migration and reversal of Japan in the years and months preceding the Tohoku-oki earthquake. a, Table of velocities (East component) within non-overlapping rectangular regions before the Tohoku-oki earthquake. Velocity for each station within each rectangular region is detrended relative to the median velocity of that station between 1 January 2006 and 8 March 2011. The medians of these detrended velocities are then chosen as the representative regional velocities. The *y* axis indicates distance in the northeast direction for the centroid of each rectangular region of Fig. 1 (beginning in the southwest). Green lines indicate the along-strike locations and times of earthquakes exceeding moment magnitude 6. **b**, A zoom-in of panel **a**, between the beginning of September 2010 and the beginning of February 2011 (dashed vertical black lines on panel **a**). The dashed line on panel **b** indicates the velocity front that migrates across Japan from the southwest. **c**, Snapshots of median displacement for stations within 1° squares (after detrending each time series) for 10-day windows from October 2010 until the Tohoku-oki earthquake (see also Supplementary Video 4, which covers this period). Green arrows indicate horizontal displacements and coloured circles indicate vertical displacements. The period covered by snapshots is the unstable period described in this study.

within the extent of the mainshock source area, whereas seismic tremor is reported as having begun in late January 2011 within 100 km of the Tohoku-oki mainshock hypocentre[23,24]. The stress field in the crust above the Tohoku-oki mainshock slip area, as determined by analysis of focal mechanisms[25], shows a gradual reduction in compression over three years in the run-up to the mainshock, although the time resolution of this reduction is limited by the number of large enough earthquakes.

Previous studies in which pre-Tohoku-oki surface trajectories were modelled[10,11,26] suggest a gradual acceleration in uncoupling of the plate interface near the mainshock region over the decade preceding the mainshock, although these analyses have not extended far beyond the mainshock rupture extents. There are no previous reports of precursory continental surface motions preceding the Maule earthquake. Although able to capture the main decadal-scale features, previous efforts to characterize secular tectonic velocity changes leading up to the Tohoku-oki and Maule earthquakes have been hindered by the need for iterative, manual improvements of the trajectory model. With GrAtSiD, we have been able to identify the subtle transients in the time series across the whole Japanese and South American networks that would have taken an unreasonable amount of time without an automated approach.

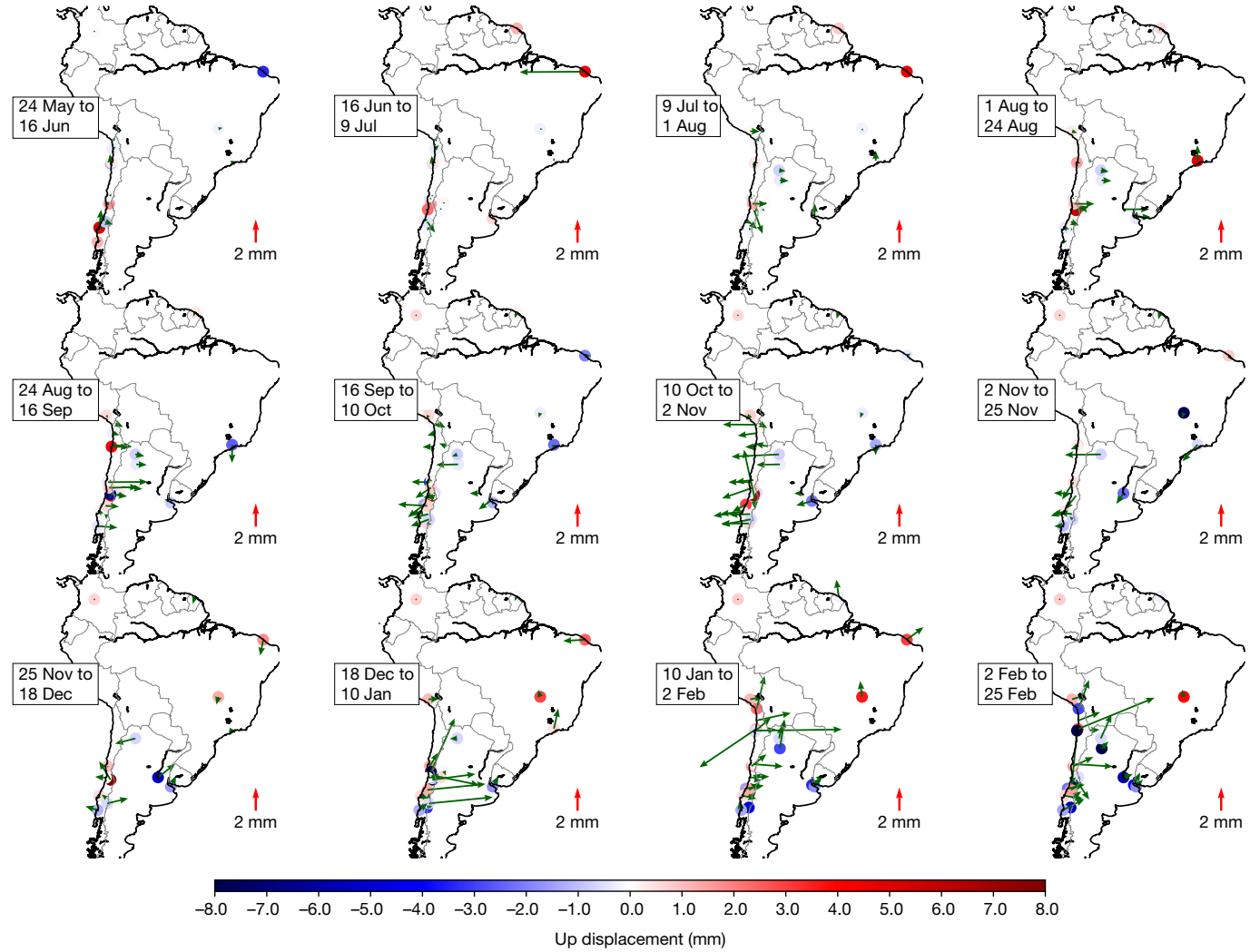


Fig. 4 | Visualizing the reversals motions in South America in the months preceding the Maule earthquake. Snapshots of displacement (after detrending) for 23-day windows from 279 days until 2 days before the Maule Earthquake. This time period covers the unstable period described in our study. Green arrows indicate horizontal displacement and coloured circles indicate vertical displacement.

Investigation of deformation mechanisms

To investigate the physical processes responsible for the pre-Tohoku-oki unstable period, we modelled the displacements (obtained from the trajectory modelling) using a simple elastic dislocation approach[27,28] (Methods). The model supports the possibility of an initial migration of slip across the subduction interface and indicates that the unstable period began with a large slab extension localized in the downgoing Philippine Sea Plate (Fig. 5 and Supplementary Video 7). This proposed slow slip across the margin is apparent from the general increase in apparent slip along the fault model from southwest to northeast. The occurrence of slow slip is also supported by similar propagation velocities (a few kilometres per hour) of the signal along the continental surface that have been observed in the Cascadia subduction zone[29] and the Nankai Trough[30]. Furthermore, the velocities of the in situ slip (of the order of 1 mm per day) match the velocities expected for slow-slip events[31]. The extension appears in the simplified dislocation model as an apparent simultaneous shallow updip slip and deep downdip locking (backslip[32]) centred at approximately 50–100 km. Following this extension, the surface velocity swings between pointing at the subduction trench, towards Eurasia, and then again towards the subduction trench. During this swing, the elastic dislocation model seems to indicate that the plate interface switches between enhanced locking and creeping: a more likely explanation is that the model is incorrectly projecting deformation of a viscoelastic rebound to the sudden extension. Figure 6 illustrates the possible series of events in the pre-Tohoku-oki unstable period. The initial slow-slip migration from the southwest to northeast of the study region is possibly facilitated by fluid release onto the plate interface. Seismic tomography of the downgoing Philippine Sea Plate[33] shows Poisson's ratio values that indicate a slab with fluid-filled porosity at the depths where we suspect slow slip to have begun and where extension is thought to be centred. Tomography of the Pacific slab also indicates sufficient fluids from the depth of the continental Mohorovičić discontinuity (Moho)[34] and shallower[35]. Such an expulsion of fluids and propagation of the slow slip would be supported by geological evidence for chemical mineral reactions. These reactions, that typically last between one and four months[36], produce pulse-like slab dehydrations that in turn suddenly increase pore-fluid pressure, thereby causing a feedback effect that mechanically weakens the plate interface[31,36,37]. This timescale, taken as an upper bound on the duration of these processes, fits the timescale of the suspected slow-slip duration that we describe in this study.

We hypothesize that the sudden extension that began towards the end of October could have been caused by either eclogitization of the slab just below the extension centre (for example, 50–100 km), or deeper slab pulling. Eclogitization processes can be rapid and mostly aseismic under thermodynamically overstepped conditions[36,38] in which pressure and

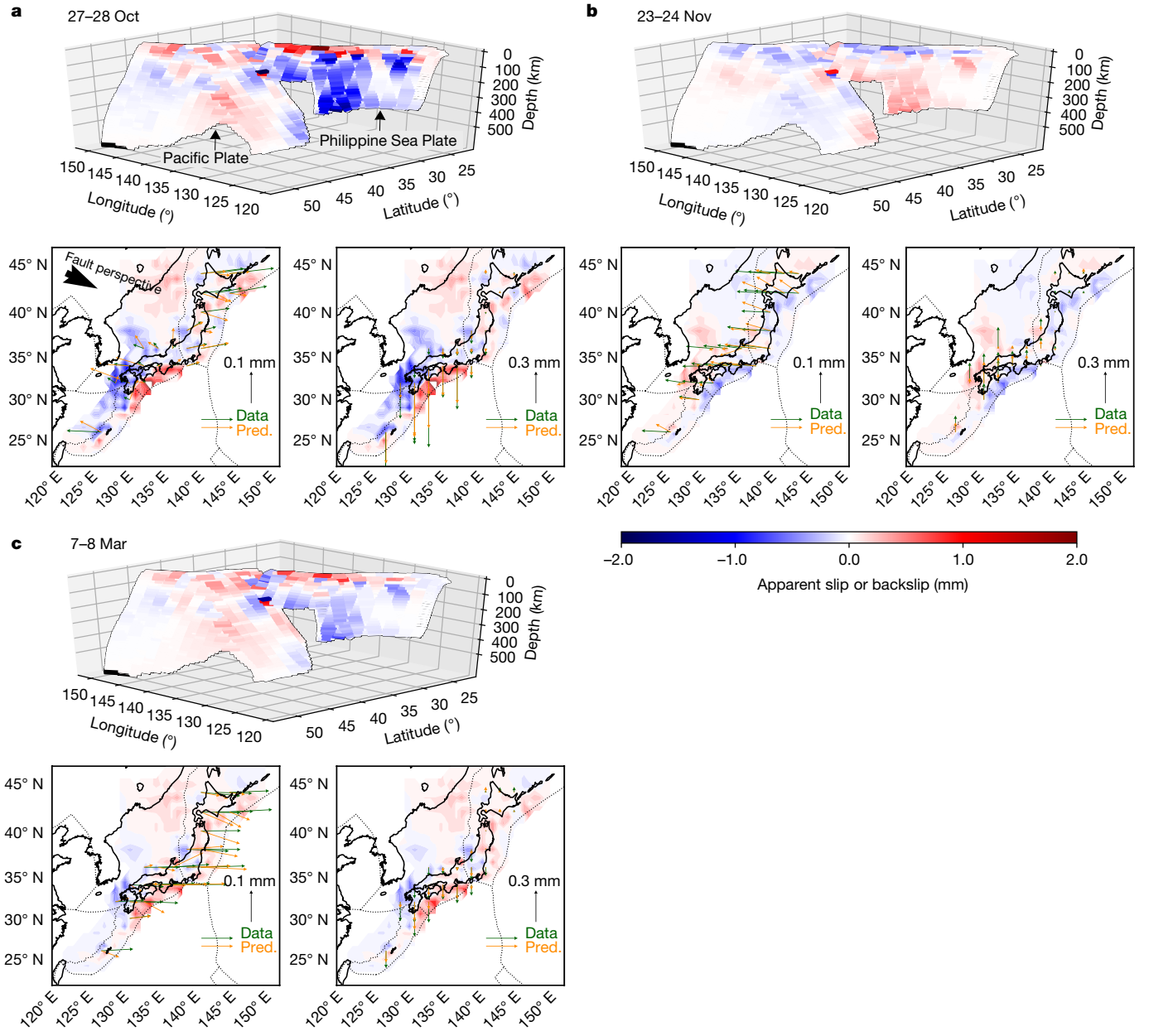


Fig. 5 | Surface velocities and kinematic models of apparent slip or backslip for three daily time windows during the unstable period preceding the Tohoku-oki earthquake. The date of the daily model is indicated above the view of the fault (all frames from September 2010 onwards are shown in Supplementary Video 8). Corresponding data and model predictions (for horizontal and vertical displacements) are shown in the maps underneath each view of the fault. The fault consists of two downgoing oceanic slabs in contact with the overriding continental plate: the Pacific Plate to the northeast and the Philippine Sea Plate to the southwest. On 27 October (**a**), the Philippine Sea Plate seems to be in a state of extension centred at approximately 50 km depth, as indicated by the updip positive and downdip negative dislocation pattern. Comparing the modelled kinematics of 23 November (**b**) and 7 March (**c**) (top right and bottom left subsets) we can see that the sense of dislocation is reversed for data displacements that are also approximately reversed. The junction between the two downgoing slabs also seems to demarcate a change in kinematics throughout all of the unstable phase. We note that the oscillating signal following the sudden pulldown (extension) is probably not well modelled by elastic dislocation. The apparent switch between enhanced locking, creeping and locking is likely to be an artefact of the elastic dislocation approach, which models the suspected viscoelastic response of the subduction zone as a purely plate interface process.

temperature make the reaction energetically favourable but sufficient activation energy for the initial reaction has been lacking. Conditions for fast eclogitization would require metastable regions in the downgoing slab, which receive infiltration of external fluid to overcome sluggish kinetics[39,40]. Although the presence of fluids—a catalyst for such a sudden densification to occur—is highly likely[33], the volume of metastable slab material (lower-crustal gabbroic rocks) in the Philippine Sea slab at depths of approximately 50–100 km is not known and, owing to the rather slow spreading of the related ridge (about 4 cm yr^{-1})[41], is expected to be small[42].

An alternative to pull from fast eclogitization is that the sudden pull comes from a deeper source. From the dislocation modelling, it is not clear if this initial downward slab pull has a shallow or deep origin. Figure 5 and Supplementary Video 7 show that the largest values of apparent enhanced locking are usually in the deepest patches of the model (approximately 400 km) on the Philippine Sea Plate interface. A possible

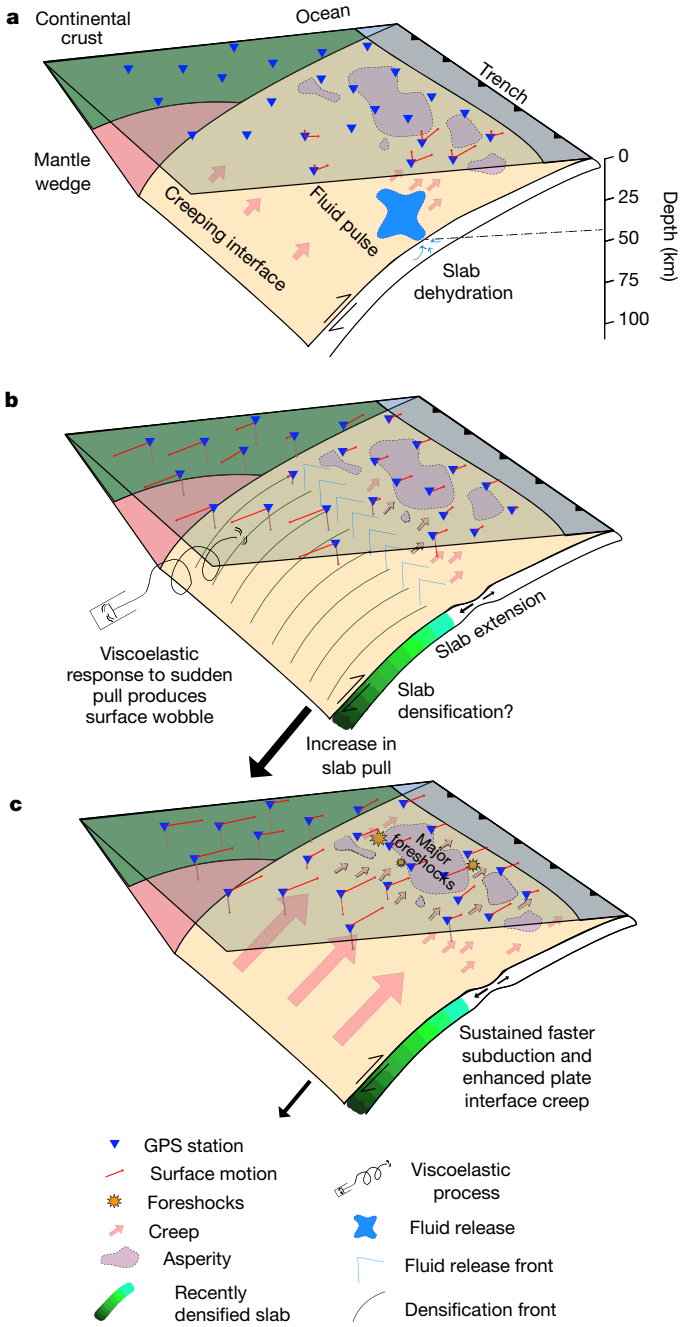


Fig. 6 | Cartoon (not to scale) to illustrate the possible processes explaining the deformation recorded in the unstable period captured before the Tohoku-oki earthquake. a, An initial pulse of fluids is released onto the plate interface via intraslab fluid pathways and dehydration reactions. This facilitates a slow-slip event caused by a reduction in the effective coefficient of friction. **b**, Slow-slip front migrates along strike. Sudden slab densification occurs and creates a buoyancy instability and enhanced downwards pull. This initial downwards acceleration causes slab extension and a quasidynamic wobbling of the subduction zone as the slab viscoelastically rebounds. Slab extension also results in faster creep downdip of the seismogenic zone and below the continental Moho. **c**, The slab extension continues, resulting in prolonged deep slip. The sudden slab pull and viscoelastic response, along with the large slow slip facilitated by fluid release has resulted in enhanced creep in the seismogenic zone that critically stresses a configuration of mature asperities, resulting in large foreshocks and the mainshock Tohoku-Oki M_w 9.0 earthquake.

mechanism to explain this deeper pull would be buoyancy instability caused by a sudden densification of slab material from alpha-olivine into beta-olivine or even gamma-olivine (that is, wadsleyite and ringwoodite formation) as it moves through the mantle transition zone at depths of 410 km and 660 km (ref. [43]). Intrinsic chemical heterogeneities would cause the reaction to start at spatially heterogeneously distributed locations within the slab; the reaction would then spread with increasing velocity owing to the exothermic reaction and resulting densification[43]. The sudden increase of slab pull would cause a perturbation of the force equilibrium at the plate interface sufficient to trigger a nonlinear response such as shear-heating-related slip acceleration or a power-law stress-dependent creep[44], which by its nature is a transient phenomenon.

From the tomography[45], the Philippine Sea Plate at mantle transition-zone depths seems to be in contact with (or in close proximity to) the slab of the metastable Pacific Plate[46], which has subducted under the Philippine Sea Plate in a flat fashion[45]. Therefore, it could be the sudden densification of metastable material in the Pacific slab and its subsequent dragging against the Philippine Sea slab that has caused the increased slab pull. Related to this possibility of slab communication, the dislocation model (Fig. 5 and Supplementary Video 7) suggests an interaction between downgoing slabs at their lateral juncture—whereby it seems that the plunging Philippine Sea Plate has dragged the Pacific Plate, which also appears to undergo some extension in the unstable period. This slab communication, however, is only apparent from the simple dislocation model, which has projected all deformation onto the fault plane. Accordingly, future work should focus on clarifying the roles of each slab in the observed unstable period, with the inclusion of processes additional to elastic dislocation.

This sudden extension in the slab before the Tohoku-oki earthquake that we describe in this work was previously reported in a study of satellite gravity measurements[47], although the statistical significance of the reported gravity gradient anomaly is disputed[48]. This anomalous gravity-gradient signal begins a month later than the onset of extension recorded by GNSS in our study. Assuming that the reported gravity-gradient anomaly is valid, then a delay of just one month in these observations would indicate that these independent measurements are capturing the same process. The discrepancy in onset could also be reduced if we consider monthly sampling of the gravity data. The favoured extension model for the gravity gradient signal is a deep normal faulting in the Pacific Plate at depths greater than 245 km, whereas our modelling suggests that extension (the hinge line between apparent enhanced slip and locking) is shallower, at depths of approximately 50–100 km and with the extension signal starting in the Philippine Sea Plate.

Compared to the pre-Tohoku-oki case, prior to the Maule earthquake there were far fewer GNSS stations and much higher variability in station spacing. Therefore, elastic dislocation modelling for the Maule unstable period was not carried out. Nevertheless, owing to the reversing nature of the pre-Maule signal, we suspect that a sequence of events consisting of deeper slab pull interacting with bursts of slow slip further updip—similar to what occurred before the Tohoku-oki earthquake—also occurred before the Maule earthquake (see discussion in the Methods).

A spectrum of interseismic processes

In conclusion, we have identified enhanced slab pulling as the likely driving mechanism for the wobbles observed before the Tohoku-oki and Maule earthquakes. However, had the asperities updip of the pulling not been mature, it is conceivable that the unstable periods caused by enhanced slab pulling would have ceased without any great earthquake occurring. Likewise, we suggest that overdue (mature) earthquake segments might only tend to rupture after receiving sufficient additional shear stress from an acceleration in deeper subduction. In this context, the earthquake research community should be considering the frictional conditions of the seismogenic plate interface in

conjunction with the variable boundary conditions that seem to have a role in bringing large mature segments to failure.

Comparing the timing of the unstable phase reported in our study to the reported foreshock activity of other studies is not straightforward, owing to considerations of seismic network resolution and varied waveform processing strategies. Although there is anomalous seismicity identified before the Tohoku-oki and Maule earthquakes on the timescale of one to a few months before the mainshock, most of this analysis has been focused closer to the mainshock location. The large wobbles reported in our study appear to be mainly aseismic and affect far-field regions larger than the extent of the rupture zone on the subduction plate interface. It is worth noting here that the adjective 'aseismic' is defined by the threshold for which events can be detected with existing seismological networks. This threshold can be limited both by seismic processing strategies and network geometry. Therefore, there may be some as-yet-undetected seismic signals either along the plate interface or deeper inside the slab that correlate with these large-scale anomalous GNSS recorded motions.

Finally, we have focused on describing the most striking transient tectonic motions that we observe before the Maule and Tohoku-oki earthquakes. There are, however, many other tectonically intriguing signals in the periods analysed (Supplementary Videos 3 and 5), some of which could be related to other, smaller-magnitude earthquakes. Although we are able to notice large, tectonically driven wobbles of the surface before the Maule and Tohoku-oki earthquakes, we still do not know how prevalent these wobbles are at plate boundary zones worldwide, whether they scale according to subsequent earthquake magnitude, and therefore whether or not they can be used as reliable precursors for the imminence of major earthquakes. What is clear from this study is that subduction zones are highly dynamic on the human observable timescale and that whereas some transient events are apparent on the thousand-kilometre scale, others are very localized. To improve our understanding of the interaction of boundary conditions on relatively smaller potential rupture zones it will be essential to harness the growing amounts of GNSS data[49] to investigate the interaction of tectonic systems on the continental scale.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2212-1.

1. Kajitani, Y., Chang, S. E. & Tatano, H. Economic impacts of the 2011 Tohoku-Oki earthquake and tsunami. *Earthq. Spectra* **29**, 457–478 (2013).
2. Sagiya, T., Miyazaki, S. I. & Tada, T. Continuous GPS array and present-day crustal deformation of Japan. *Pure Appl. Geophys.* **157**, 2303–2322 (2000).
3. Báez, J. C. et al. The Chilean GNSS network: current status and progress toward early warning applications. *Seismol. Res. Lett.* **89**, 1546–1554 (2018).
4. Heki, K. & Mitsui, Y. Accelerated Pacific plate subduction following interplate thrust earthquakes at the Japan trench. *Earth Planet. Sci. Lett.* **363**, 44–49 (2013).
5. Loveless, J. P. & Meade, B. J. Two decades of spatiotemporal variations in subduction zone coupling offshore Japan. *Earth Planet. Sci. Lett.* **436**, 19–30 (2016).
6. Melnick, D. et al. The super-interseismic phase of the megathrust earthquake cycle in Chile. *Geophys. Res. Lett.* **44**, 784–791 (2017).
7. Bedford, J. & Bevis, M. Greedy automatic signal decomposition and its application to daily GPS time series. *J. Geophys. Res. Solid Earth* **123**, 6992–7003 (2018).
8. Reid, H. F. *The Mechanics of the Earthquake, The California Earthquake of April 18, 1906, Report of the State Investigation Commission* Vol. 2, 16–28 (Carnegie Institution of Washington, 1910).
9. Schurr, B. et al. Gradual unlocking of plate boundary controlled initiation of the 2014 Iquique earthquake. *Nature* **512**, 299–302 (2014).
10. Mavrommatis, A. P., Segall, P. & Johnson, K. M. A decadal-scale deformation transient prior to the 2011 Mw 9.0 Tohoku-oki earthquake. *Geophys. Res. Lett.* **41**, 4486–4494 (2014).
11. Wang, K. et al. Learning from crustal deformation associated with the M9 2011 Tohoku-oki earthquake. *Geosphere* **14**, 552–571 (2018).
12. Kato, A. et al. Propagation of slow slip leading up to the 2011 Mw 9.0 Tohoku-Oki earthquake. *Science* **335**, 705–708 (2012).
13. Johnson, P. A. et al. Acoustic emission and microslip precursors to stick-slip failure in sheared granular material. *Geophys. Res. Lett.* **40**, 5627–5631 (2013).
14. Kaproth, B. M. & Marone, C. Slow earthquakes, preseismic velocity changes, and the origin of slow frictional stick-slip. *Science* **341**, 1229–1232 (2013).
15. Lohman, R. B. & Murray, J. R. The SCEC geodetic transient detection validation exercise. *Seismol. Res. Lett.* **84**, 419–425 (2013).
16. Dill, R. & Dobslaw, H. Numerical simulations of global-scale high-resolution hydrological crustal deformations. *J. Geophys. Res. Solid Earth* **118**, 5008–5017 (2013).
17. van Dam, T. et al. Crustal displacements due to continental water loading. *Geophys. Res. Lett.* **28**, 651–654 (2001).
18. Heki, K. Snow load and seasonal variation of earthquake occurrence in Japan. *Earth Planet. Sci. Lett.* **207**, 159–164 (2003).
19. Wdowinski, S., Bock, Y., Zhang, J., Fang, P. & Genrich, J. Southern California permanent GPS geodetic array: spatial filtering of daily positions for estimating coseismic and postseismic displacements induced by the 1992 Landers earthquake. *J. Geophys. Res. Solid Earth* **102**, 18057–18070 (1997).
20. Bouchon, M. et al. Potential slab deformation and plunge prior to the Tohoku, Iquique and Maule earthquakes. *Nat. Geosci.* **9**, 380–383 (2016).
21. Gardonio, B. et al. Seismic activity preceding the 2011 M w 9.0 Tohoku earthquake, Japan, analyzed with multidimensional template matching. *J. Geophys. Res. Solid Earth* **124**, 6815–6831 (2019).
22. Ito, Y. et al. Episodic slow slip events in the Japan subduction zone before the 2011 Tohoku-Oki earthquake. *Tectonophysics* **600**, 14–26 (2013).
23. Ito, Y., Hino, R., Suzuki, S. & Kaneda, Y. Episodic tremor and slip near the Japan Trench prior to the 2011 Tohoku-Oki earthquake. *Geophys. Res. Lett.* **42**, 1725–1731 (2015).
24. Katakami, S. et al. Spatiotemporal variation of tectonic tremor activity before the Tohoku-Oki earthquake. *J. Geophys. Res. Solid Earth* **123**, 9676–9688 (2018).
25. Becker, T. W., Hashima, A., Freed, A. M. & Sato, H. Stress change before and after the 2011 M9 Tohoku-oki earthquake. *Earth Planet. Sci. Lett.* **504**, 174–184 (2018).
26. Yokota, Y. & Koketsu, K. A very long-term transient event preceding the 2011 Tohoku earthquake. *Nat. Commun.* **6**, 5934 (2015).
27. Pollitz, F. F. Coseismic deformation from earthquake faulting on a layered spherical Earth. *Geophys. J. Int.* **125**, 1–14 (1996).
28. Hayes, G. P. et al. Slab2, a comprehensive subduction zone geometry model. *Science* **362**, 58–61 (2018).
29. Rogers, G. & Dragert, H. Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. *Science* **300**, 1942–1943 (2003).
30. Yamashita, Y. et al. Migrating tremor off southern Kyushu as evidence for slow slip of a shallow subduction interface. *Science* **348**, 676–679 (2015).
31. Ide, S., Beroza, G. C., Shelly, D. R. & Uchide, T. A scaling law for slow earthquakes. *Nature* **447**, 76–79 (2007).
32. Savage, J. C. A dislocation model of strain accumulation and release at a subduction zone. *J. Geophys. Res. Solid Earth* **88**, 4984–4996 (1983).
33. Kodaira, S. et al. High pore fluid pressure may cause silent slip in the Nankai Trough. *Science* **304**, 1295–1298 (2004).
34. Tsuji, Y., Nakajima, J. & Hasegawa, A. Tomographic evidence for hydrated oceanic crust of the Pacific slab beneath northeastern Japan: implications for water transportation in subduction zones. *Geophys. Res. Lett.* **35**, L14308 (2008).
35. Liu, X., Zhao, D. & Li, S. Seismic attenuation tomography of the Northeast Japan arc: insight into the 2011 Tohoku earthquake (Mw 9.0) and subduction dynamics. *J. Geophys. Res. Solid Earth* **119**, 1094–1118 (2014).
36. Taetz, S., John, T., Bröcker, M., Spandler, C. & Stracke, A. Fast intraslab fluid-flow events linked to pulses of high pore fluid pressure at the subducted plate interface. *Earth Planet. Sci. Lett.* **482**, 33–43 (2018).
37. Ujiie, K. Chemical origin of tectonic tremor. *Nat. Geosci.* **12**, 962–963 (2019).
38. Incel, S. et al. Reaction-induced embrittlement of the lower continental crust. *Geology* **47**, 235–238 (2019).
39. Austrheim, H. Eclogitization of lower crustal granulites by fluid migration through shear zones. *Earth Planet. Sci. Lett.* **81**, 221–232 (1987).
40. John, T. & Schenk, V. Partial eclogitisation of gabbroic rocks in a late Precambrian subduction zone (Zambia): prograde metamorphism triggered by fluid infiltration. *Contrib. Mineral. Petrol.* **146**, 174–191 (2003).
41. Seno, T. & Maruyama, S. Paleogeographic reconstruction and origin of the Philippine Sea. *Tectonophysics* **102**, 53–84 (1984).
42. Bach, W. & Früh-Green, G. L. Alteration of the oceanic lithosphere and implications for seafloor processes. *Elements* **6**, 173–178 (2010).
43. Burnley, P. C., Green, H. W. & Prior, D. J. Faulting associated with the olivine to spinel transformation in Mg_2GeO_4 and its implications for deep-focus earthquakes. *J. Geophys. Res. Solid Earth* **96**, 425–443 (1991).
44. Thielmann, M., Rozel, A., Kaus, B. J. P. & Ricard, Y. Intermediate-depth earthquake generation and shear zone formation caused by grain size reduction and shear heating. *Geology* **43**, 791–794 (2015).
45. Fukao, Y. & Obayashi, M. Subducted slabs stagnant above, penetrating through, and trapped below the 660 km discontinuity. *J. Geophys. Res. Solid Earth* **118**, 5920–5938 (2013).
46. Kawakatsu, H. & Yoshioka, S. Metastable olivine wedge and deep dry cold slab beneath southwest Japan. *Earth Planet. Sci. Lett.* **303**, 1–10 (2011).
47. Panet, I., Bonvalot, S., Narteau, C., Remy, D. & Lemoine, J. M. Migrating pattern of deformation prior to the Tohoku-Oki earthquake revealed by GRACE data. *Nat. Geosci.* **11**, 367–373 (2018).
48. Wang, L. & Burgmann, R. Statistical significance of precursory gravity changes before the 2011 M_w 9.0 Tohoku-Oki earthquake. *Geophys. Res. Lett.* **46**, 7323–7332 (2019).
49. Blewitt, G., Hammond, W. C. & Kreemer, C. Harnessing the GPS data explosion for interdisciplinary science. *Eos* **99**, https://doi.org/10.1029/2018EO104623 (2018).

Methods

GNSS time series

For the South American network, all GNSS data were organized in units of 24-h periods and were processed using the Earth Parameter and Orbit System software (EPOS)[50]. The data processing was done in three steps. In the first step, all the stations were processed in a precise point positioning (PPP) model using the GNSS satellite clock and orbit products of the GFZ German Research Centre for Geosciences[51]. During PPP processing, bad observations and outliers were removed. In the second step, we followed a network-processing strategy based upon the PPP solutions and remaining observations. Up to 56 well distributed International GNSS Service (IGS) core stations are included in the network. Since the number of project stations can reach up to about 800, and the Earth Parameter and Orbit System software can process up to 250 stations for one single network, the project stations were processed in several sub-networks. Depending on data availability, ten well distributed IGS stations were included in each sub-network to connect all the sub-networks. In the network processing, all the stations were treated with the same weight. Since we did not fix any station in the network processing, the coordinate solution is the same as a free network solution. The datum of the coordinate solution was defined by the satellite orbit and clock products used in the 2008 realization of the International Terrestrial Reference Frame (ITRF2008; http://itrf.ensg.ign.fr/). In the third and final step, the network solution was aligned to the IGS second reprocessing combined daily coordinate product in ITRF2014[52,53] to reduce the impact of the datum effect. To avoid the artefacts caused by reference stations, the IGS second solutions were visually inspected. Those IGS station time series that show transient signals in the unstable period defined in this study were not used for the alignment. The coordinate results from the network solution have an accuracy of a few millimetres. IGS stations used in the network data processing are indicated in Extended Data Fig. 1.

For the Japanese case, we use coordinates from the network solutions (version F3), provided by the Geospatial Information Authority of Japan[54].

GNSS data postprocessing for the pre-Maule-earthquake solutions

We considered time series between 1 January 2005 and 25 February 2010. Time series were retained if they had a minimum of four years of data and also spanned the year preceding the Maule earthquake. Obvious, grossly outlying points were removed from the network solutions by manual inspection. Likewise, any very large artificial steps were removed. Data were further de-spiked by the application of a median filter and removal of points far from this filtered trace. We note that the median filter is only used for outlier detection and the unfiltered series, with outliers removed, is fed into the GrAtSiD algorithm.

GNSS data postprocessing for the pre-Tohoku-oki-earthquake solutions

Time series were truncated between 1 January 2006 and 8 March 2011. Time series were retained if they had a minimum of four years of data and also spanned the year preceding the Tohoku-oki earthquake. Data were provided in degrees and so these were converted into metres. Next, the median filter approach for outlier removal (as done for South American data) was applied.

Signal decomposition with GrAtSiD

The post-processing procedure described in this section was applied to both the pre-Maule and pre-Tohoku datasets. For each directional component (East, North and Up) of each GNSS station, we applied an initial GrAtSiD fit[7] with one convergence only. GrAtSiD is a greedy linear regression routine that aims to find a minimum number of transient (multi-transient) and step (Heaviside) basis functions present in the time series. In general, greedy algorithms sequentially test for the improvement of fit when including additional basis functions in the model[55]. Basis functions that are always present in the GrAtSiD solution (permanent functions) are the Fourier functions for annual and semi-annual oscillations, the first-order polynomial terms (linear trend and constant shift), and the steps imposed owing to known hardware changes or when the station is within a radial cut-off distance from a catalogued earthquake, with the cut-off distance r (in kilometres), being $r = 10^{0.5M-0.8}$, where M is the earthquake magnitude (as also used by the Nevada Geodetic Laboratory[49]). For the Maule data, we considered all events with at least magnitude 5, as provided by the catalogue of the National Earthquake Information Center of the United States Geological Survey (USGS NEIC) (https://earthquake.usgs.gov/earthquakes/search/), whereas for the Tohoku-oki data we considered moment magnitudes ≥6, again taken from the USGS NEIC catalogue. Permanent functions also include the known times of steps due to equipment- or processing-related discontinuities.

The regression model, $x(t)$, of GrAtSiD is formulated as follows:

$$x(t) = mt + d + \sum_{k=1}^{n_k} [s_k \sin(\omega_k t) + s_k \cos(\omega_k t)] + \sum_{j=1}^{n_j} b_j H(t - t_j) + \sum_{r=1}^{n_r} \sum_{i=1}^{n_i} [A_i (1 - e^{-(t-t_r)/T_i})] + \xi(t)$$

in which t is time, and m and d are the coefficients of the first-order polynomial. The sine and cosine functions model the annual and semi-annual background seasonal oscillations in the signals with $n_k = 2$ (coefficients s_k and c_k). $H(t - t_j)$ represent the Heaviside functions that are pre-defined as basis functions or automatically found with GrAtSiD (with b_j being the coefficients for events at time t_j). The double summation term represents the r multi-transients, which are the sum of exponential functions starting at t_r. T_i are the decay constants and we use three for each multi-transient. A multi-transient is a decaying function that is made from the sum of multiple decay functions—in this case we implement exponential decay functions but we could also use logarithmic decays. By not constraining the signs of coefficients (A_i), we can produce, with one or two multi-transients in series, a variety of signal shapes. Accordingly, the multi-transient is chosen as a sparse basis function in the regression because it is a versatile function that can capture transient signals of varying durations. $\xi(t)$ is the residual. The above equation looks very similar to the Extended Linear Trajectory Model (ELTM) described in equation (10) of ref. [56] except that the transient functions (single transients) are replaced with multi-transients. Full details of the algorithm can be found in the GrAtSiD methods paper[7] along with videos that illustrate the convergence of the solution for multiple examples.

The purpose of the initial GrAtSiD fit is to estimate a residual time series for each component of each station. Using the common-mode error reduction approach[19], we then took the median residual value for each directional component (as a function of time) and subtracted this from the corresponding time series. Next, we applied GrAtSiD again on the common-mode corrected time series. The common-mode filter also serves as a low-pass filter and is effective in removing high-frequency noise such as reference frame jitter[56] as well as reducing heteroskedasticity in the network solutions. Five convergences were run for each time series of the Japanese F3 solutions and 30 were run for the pre-Maule-earthquake time series. The solution that is most similar to other convergences was chosen as the solution for the decomposition. Similarity is defined by the solution that has the smallest average residual to all other solutions[7]. For Fig. 3, Extended Data Fig. 4, and Supplementary Videos 3–8, the seasonal contributions and the steps in the GrAtSiD trajectories were removed: that is, the Fourier functions and Heaviside functions (that have been imposed from seismic catalogues or automatically found with GrAtSiD), are removed from the solution

(simply done by removing these terms from the linear combination of basis functions).

Extended Data Figs. 2 and 3 show example results of the signal decomposition and how time series are cleaned by the removal of background seasonal oscillation and common-mode noise (see also Supplementary Videos 1 and 2). Extended Data Fig. 4 shows the velocity table for all three directional components in Japan (as shown in the East component only in Fig. 3a) while the Source Data for Fig. 3 gives the non-overlapping along-strike rectangular regions used in this analysis. Extended Data Fig. 5 shows the effect of removing modelled terms from the trajectory model that was optimized by GrAtSiD.

Investigating possible non-tectonic sources of the transient signal

The following sub sections describe the tests done to determine whether the unstable periods before the Maule and Tohoku-oki earthquakes, defined in this study by times where we observe the strong transient motions in the GNSS time series, are likely to be non-tectonic in origin. Possible non-tectonic sources that we investigate include: (1) artefacts from GNSS processing, such as systematic and sustained distortion of the reference frame over large distances, and (2) unusually strong interannual variations in seasonal fluid loading of the Earth's surface (oceanic, atmospheric and hydrological) that would be decomposed into the non-seasonal portion of the signal by the GrAtSiD algorithm (GrAtSiD assumes a steady background oscillation described by annual and semi-annual Fourier functions).

Investigation of possible processing-related artefacts. As stated above, for both sets of time series, we applied a common-mode filter that removes much of the higher-frequency reference frame jitter in the network solution. To reduce the chance of having large lower-frequency distortions of the reference frame (which can appear to be tectonic transients in the region of interest) one must exclude any stations that cannot be well modelled by the assumed trajectory model from the list of stations used to define the reference frame[56]. For the pre-Maule-earthquake analysis, we followed a careful selection of reference stations that are largely free of artefacts and, furthermore, we selected stations far away from the regions suspected to be strongly affected by tectonic transient motion. As an additional test, we split our analysed time series into two groups: group 1 we deem to be affected by the unstable period and group 2 we deem to be unaffected. Extended Data Fig. 6 shows a map and median variation in detrended velocity in each directional component for both groups. We can see that the strong east–west wobbling is most pronounced in stations that are not used in defining the reference frame and in regions where we are interpreting the pre-Maule unstable transient period to be manifest. There is a hint of some of the transient signal of group 2 in the North and Vertical component correlating with the Group 1 East component, although this analysis is hindered by the relative lack of stations in regions far away from the suspected transient-affected region. Furthermore, it can be observed in Supplementary Videos 5 and 6 that the wobble of the unstable period in the East–West direction is strongest along the Group 1 stations. The observation that the East–West unstable period transient motion is not as pronounced in Group 2 points towards the observed transient in Group 1 being largely unrelated to processing artefacts.

One notable feature of the pre-Tohoku-oki wobble that cannot be easily explained by a reference frame problem is the propagation of the wobble signal (as shown in Fig. 3a, b). Still, for the Tohoku-oki earthquake we wanted to perform an analysis similar to that done for the Maule earthquake, in which we look for the spatial extents of the wobbling signal. For the Japanese F3 network solutions, we are not able to replicate such an analysis because the whole of the network seems to be affected during the unstable period. Therefore, we took IGS08 (the 2008 datum of the IGS) PPP solutions[57,58] from the Nevada Geodetic Laboratory[49] for stations spanning a wider spatial extent on the globe. The signal decomposition is as described above (that is, outlier removal and application of GrAtSiD for signal decomposition) but without the application of a common-mode filter (since high-frequency noise in the PPP solutions correlates less on the daily scale compared to network solutions). Extended Data Fig. 7 shows the extent of the network with group 1 (affected stations) and group 2 (largely unaffected stations) with the median variation in detrended velocity of each directional component for both groups. From Extended Data Fig. 7 and the video of the unstable period for this larger-spatial-scale dataset (Supplementary Video 8) we can see that the affected region spans into the Korean Peninsula, China, Russia, and elsewhere, with a hint of extension between Japan and the Korean Peninsula and Russia. Much further away, we see that there is minimal variation in the detrended velocities during this unstable period (Supplementary Video 8).

A direct comparison of the Nevada Geodetic Laboratory (NGL)'s PPP and the Geospatial Information Authority of Japan (GSI)'s F3 solutions can be seen in Extended Data Fig. 8. Here we see that the wobble of the unstable period is apparent in both sets of solutions. From this analysis of the GNSS time series before the Tohoku-oki earthquake, we conclude that the transient of the unstable period is most probably not an artefact of GNSS processing. Furthermore, from both velocity analysis of modelled trajectories and visual comparisons of daily solutions we see the same duration and motions of the unstable period for the East component, both in the PPP and network F3 solutions. In the North and Vertical components of the PPP, we do not see the unstable period in group 1 as clearly as we do for the F3 solutions. This is partially due to the daily repeatability noise being larger in the PPP solutions.

Investigation into interannual variations in fluid loading. It is well established that hydrological, oceanic and atmospheric loading controls the seasonal oscillation observed in GNSS time series[17,59]. Accordingly, highly anomalous changes in the amount of fluid loading can greatly affect the amplitude of the oscillation of a particular season[60], whereas longer changes (for example, Earth's current rapid climate change) can manifest themselves as noticeable changes in the background trend of GNSS time series[61,62], particularly in the vertical component of motion. One simplifying assumption of the GrAtSiD algorithm is that there exists a constant background seasonal oscillation that is modelled as the sum of annual-period and semi-annual-period sine and cosine functions. Accordingly, GrAtSiD will map any strongly anomalous (unseasonal) fluid-loading-induced motion into the sparse portion of the signal that comprises multi-transient functions[7]. Such an assumption, however, is workable when we do not anticipate too much anomalous seasonal motion in the time series. Given the constraints in the assumptions of the GrAtSiD algorithm, we performed an additional investigation into whether the fluid loading (caused by precipitation, and by oceanic and atmospheric loading) could be a strong candidate to explain the unstable period motion before the Maule and Tohoku-oki earthquakes.

At all GNSS station locations, we gathered predictions of displacement due to fluid loading and decomposed these into annual and interannual surface displacements using GrAtSiD. The predictions for surface displacements are a sum of predictions provided by the GeoForschungsZentrum Potsdam Earth-System-Modelling (http://esmdata.gfz-potsdam.de:8080/repository)[16]. We summed the predictions of non-tidal atmospheric loading, non-tidal oceanic loading, hydrological loading and barystatic sea-level loading (NTAL + NTOL + HYDL + SLEL) to produce daily time series of predicted GNSS displacement in the same three directional components as the GNSS time series. All of these prediction products are generated by convolving fluid loading measurements with Green's functions of Earth's elastic response[16].

For both GNSS measured and fluid predicted displacements we applied the same processing flow: first we ran GrAtSiD to fit a trajectory model to each time series. Next, we isolated the multi-transient and polynomial terms for each time series and generated velocity time series. From these velocity time series we subtracted the median velocity of the observation period to leave behind a time series measuring the

deviation from median velocity (the deviation from the 'steady-state' velocity at each station). Finally, we take the median of these median corrected velocities across all time series to produce a measure of the overall transient motion of the stations. This is the same approach as described above (Extended Data Figs. 6, 7 and Fig. 2c, f). For the Tohoku-oki and Maule earthquakes, respectively, Extended Data Figs. 9 and 10 show the maps and time series of stations used in the comparison of measured and predicted displacements for the Tohoku-oki and Maule earthquakes, respectively.

From comparing the predicted and observed velocities in Extended Data Figs. 9 and 10, we see that there is no strong agreement in the onset of the observed unstable phase and the predicted velocities. This is especially so in the East component in both pre-Maule and pre-Tohoku-oki data, for which the strongest explanation of this motion in this component remains tectonics. Perhaps the strongest indication that fluid loading might be involved in the observed unstable period can be seen in the Vertical component in Japan. In this case, the median predicted deviation of velocity of the network is relatively unstable compared to the previous period across the whole network from mid-2010 until the onset of the Tohoku-oki earthquake and, although both the polarity of motion and onset of the predicted and observed unstable phases are not in good agreement, it could be that the unusually strong fluctuations in fluid loading (visible in the Vertical prediction) have contributed to tipping a critically stressed subduction system into instability.

Elastic dislocation modelling of the pre-Tohoku-oki unstable phase

To aid in the interpretation of the surface signal of the unstable phase, we constructed a kinematic model in which the surface displacements are predicted using Green's functions of elastic dislocation on the subduction plate interface in a layered isotropic spherical Earth[27]. The major drawback of such modelling is in the assumption that all motion can be related to plate interface kinematics in a purely elastic medium, whereas a more realistic (but harder to model) scenario includes processes such as viscoelastic and viscoplastic deformation.

From the Slab2 subduction interface geometry[28] we constructed 1,151 rectangular patches from which the Green's functions for elastic dislocation were calculated using the default Earth layering parameter file of the Static-1D software (https://github.com/fpollitz-usgs/STATIC1D)[27]. We used the convex optimization software CVXOPT[63] to solve for the updip dislocation value, m in the following problem: minimize$\{||d - Gm|| + \lambda||m||\}$, where d are the surface displacements, G is the matrix of Green's functions containing the surface predictions for unitary updip dislocations on each patch, and λ is the damping factor to regularize the inversion. We allowed the solution to find both negative and positive updip dislocation values, whereby negative values indicate transient enhanced locking[32] and positive values indicate transient slip. This freedom is given to the model because the nature of the signal (wobbling in East–West and Vertical directions) possibly indicates that there were periods of both enhanced slip and locking in this unstable phase. Furthermore, previous studies have allowed for the simultaneous presence of both enhanced locking and slip in kinematic models of the interseismic period[10,26]. The data have been interpolated onto a grid with squares of 2 decimal degrees, with each square containing at least ten stations and taking the median displacement from time series within each square, where each time series has been detrended from its median velocity. Figure 5 shows the results of the kinematic modelling at three separate points in time during the unstable phase. A separate model was made for the displacements of each day and these form the frames of Supplementary Video 7.

Interpretation of the pre-Maule unstable period signals

There are far fewer stations for the pre-Maule unstable period compared to the pre-Tohoku-oki unstable period. Nevertheless, given the similarity of the data features to those captured before the Tohoku-oki earthquake, we can hypothesize on the likely mechanisms at play. Whereas we do see some strong subsidence that could be indicative of enhanced slab pull, particularly in the back arc away from the Maule rupture zone and into Argentina, there is not a strong vertical reversal (as observed before the Tohoku-oki earthquake) and the wobbling that defines the unstable period is only clear in the East–West direction. Furthermore, there are insufficient stations to determine whether subsidence signal began with, before or after the East–West wobbling. Within the pre-Maule unstable period, there also does not appear to be a period of obvious extension. Therefore, we hypothesize that the simplest sequence of events starts with an increase in slab pull force causing a plunge (without any noticeable extension) of the downgoing slab and simultaneous eastwards motion as the locked shallower portion of the plate interface drags the continent at a higher-than-usual rate. This is followed by a large slow-slip event triggered by the increase in plate interface shear stress, which causes the overriding plate to move westwards. Fluids on this slab segment (which could facilitate the slow slip) have been inferred from tomography and are also thought to control longer-term subduction kinematics[64]. This slow slip is then arrested, while the faster velocity of the slab persists, resulting again in faster-than-usual westward motion. The next slow slip is either so close to the mainshock time that we do not resolve it before the mainshock, or it does not occur at all, with the enhanced shear stress on the plate interface causing the mainshock asperities to fail with minimal foreshock or foreslip activity. It is possible that the large slow slip that occurs during this sequence of events acts as the destabilizing event that enhances the slab pull. It is not clear how much of a viscoelastic component there is to the wobble signals recorded during the pre-Maule unstable period, although future modelling could tackle this question. The pull of the pre-Maule sequence seems to be coming from a very deep source: if we project vertically from the stations of strong subsidence down to the plate interface, we arrive at depths of several hundreds of kilometres, with depths of approximately 500 km estimated for the stations in Argentina[65], leading again to the interpretation of a buoyancy instability caused by a sudden transition of slab mantle material as it moves through a mantle transition zone.

Spatial averaging, time series detrending, and deviations from background velocity

In Fig. 3 and Extended Data Fig. 4, we plot the regional median deviation from background steady velocity, where the steady velocity of each station is first calculated by taking the median of the trend (minus seasonal and steps). Median velocity values are convenient to calculate owing to the smooth interseismic trends (the sum of the first-order polynomial and multi-transient functions) recovered from GrAtSiD. For each rectangular region, we calculate the median deviation from the background steady velocity (as a function of time) across each velocity time series in that region. Bounds for the non-overlapping regions are provided in the Source Data for Fig. 3 and Extended Data Fig. 4.

As done in Fig. 3 and Extended Data Fig. 4, we plot the median deviation in velocity from the background (also using median) velocity across groups of stations in Fig. 2c, f and Extended Data Figs. 6, 7, 9 and 10. Essentially, this allows us to generate measures of velocity that deviate from a steady state for particular directional components across groups of stations. As mentioned in the legend of Fig. 2, this average deviation from background velocity is only calculated if more than 55% of stations in the group have data on that particular day. This is to avoid spurious apparent deviations from velocity that arise from fewer samples.

For all Supplementary Videos 3–8 and for Figs. 3c and 4, the displacements are calculated from detrended time series, where the trend of the time series is determined by the median velocity of the smooth interseismic trends. Then within each regional window (for example, 2° by 2° for Supplementary Videos 3 and 4) the median detrended displacements are plotted in each frame.

Data availability

The daily GNSS displacement time series and the predicted displacements from fluid loading models are available in a data supplement to this paper[66].

Code availability

To create the maps in the figures and Supplementary videos, we used Python package Matplotlib[67] and Generic Mapping Tools[68]. The GrAtSiD code used for the trajectory modelling in this study can be provided upon request from the corresponding author.

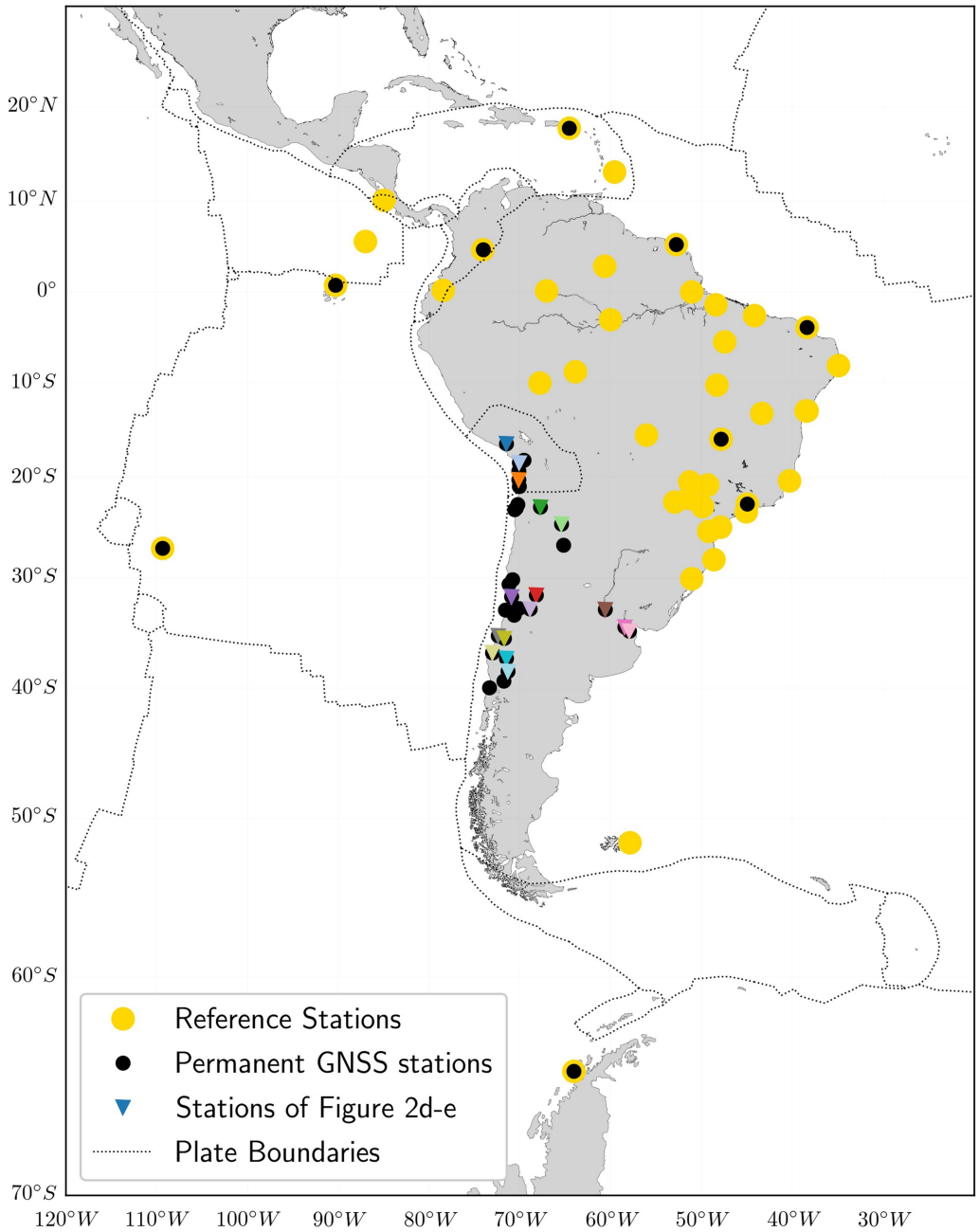
50. Gendt, G. et al. *GFZ Analysis Center of IGS. Annual Report for 2013* 1–10 (GFZ German Research Centre for Geosciences, 2013); ftp://ftp.gfz-potsdam.de/GNSS/DOCS/IGS_repro2/GFZ_igs_annual_report_2013_iss1-0.pdf.
51. Deng, Z., Fritsche, M., Nischan, T. & Bradke, M. *Multi-GNSS Ultra Rapid Orbit-, Clock- & EOP-Product Series* (GFZ Data Services, 2016); https://doi.org/10.5880/GFZ.1.1.2016.003.
52. Rebischung, P., Altamimi, Z., Ray, J. & Garayt, B. The IGS contribution to ITRF2014. *J. Geodyn*. **90**, 611–630 (2016).
53. Altamimi, Z., Rebischung, P., Métivier, L. & Collilieux, X. ITRF2014: A new release of the International Terrestrial Reference Frame modeling nonlinear station motions. *J. Geophys. Res. Solid Earth* **121**, 6109–6131 (2016).
54. Nakagawa, H., et al. *Development and validation of GEONET New Analysis Strategy (Version 4) [in Japanese] Annual Report of the Geographical Survey Institute* Vol. 118, 1–8 (GSI, 2009); https://www.gsi.go.jp/common/000054716.pdf.
55. Needell, D., Tropp, J. & Vershynin, R. Greedy signal recovery review. In *2008 42nd Asilomar Conf. on Signals, Systems and Computers* 1048–1050, https://core.ac.uk/reader/23798095 (IEEE, 2008).
56. Bevis, M. & Brown, A. Trajectory models and reference frames for crustal motion geodesy. *J. Geod*. **88**, 283–311 (2014).
57. Rebischung, P. et al. IGS08: the IGS realization of ITRF2008. *GPS Solutions* **16**(4), 483–494 (2012).
58. Zumberge, J. F., Heflin, M. B., Jefferson, D. C., Watkins, M. M. & Webb, F. H. Precise point positioning for the efficient and robust analysis of GPS data from large networks. *J. Geophys. Res. Solid Earth* **102**, 5005–5017 (1997).
59. Liu, L., Khan, S. A., van Dam, T., Ma, J. H. Y. & Bevis, M. Annual variations in GPS-measured vertical displacements near Upernavik Isstrøm (Greenland) and contributions from surface mass loading. *J. Geophys. Res. Solid Earth* **122**, 677–691 (2017).
60. Kusche, J. E. J. O. & Schrama, E. J. O. Surface mass redistribution inversion from global GPS deformation and Gravity Recovery and Climate Experiment (GRACE) gravity data. *J. Geophys. Res. Solid Earth* **110**, B09409 (2005).
61. Borsa, A. A., Agnew, D. C. & Cayan, D. R. Ongoing drought-induced uplift in the western United States. *Science* **345**, 1587–1590 (2014).
62. Bevis, M. et al. Accelerating changes in ice mass within Greenland, and the ice sheet's sensitivity to atmospheric forcing. *Proc. Natl Acad. Sci. USA* **116**, 1934–1939 (2019).
63. Grant, M. & Boyd, S. CVX: Matlab software for disciplined convex programming. Version 2.1, http://cvxr.com/cvx/citing/ (2014).
64. Moreno, M. et al. Locking of the Chile subduction zone controlled by fluid pressure before the 2010 earthquake. *Nat. Geosci*. **7**, 292–296 (2014).
65. Chen, Y. W., Wu, J. & Suppe, J. Southward propagation of Nazca subduction along the Andes. *Nature* **565**, 441–447 (2019).
66. Bedford, J. et al. Trajectory models for daily displacement time series in the five years preceding the 2010 Maule Mw 8.8, Chile, and 2011 Tohoku-oki Mw 9.0, Japan earthquakes (GFZ Data Services, 2020); https://doi.org/10.5880/GFZ.4.1.2020.001.
67. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng*. **9**, 90–95 (2007).
68. Wessel, P., Smith, W. H., Scharroo, R., Luis, J. & Wobbe, F. Generic mapping tools: improved version released. *Eos* **94**, 409–410 (2013).

Acknowledgements We thank the Geospatial Information Authority of Japan (GSI) and the Nevada Geodetic Laboratory (NGL), University of Nevada, for their assistance and for providing time series for this study. We thank Y. Bock and K. Heki for comments. J.R.B. thanks S. Sobolev for his comments. J.R.B. thanks the German Science Foundation (DFG) for grant MO-2310/3. M.M. acknowledges support from FONDECYT 1181479, the Millennium Nucleus "The Seismic Cycle Along Subduction Zones" grant NC160025, and the Research Center for Integrated Disaster Risk Management (CIGIDEN), CONICYT/FONDAP 15110017. J.C.B. acknowledges support from FONDECYT projects 1170430 and 1181479.

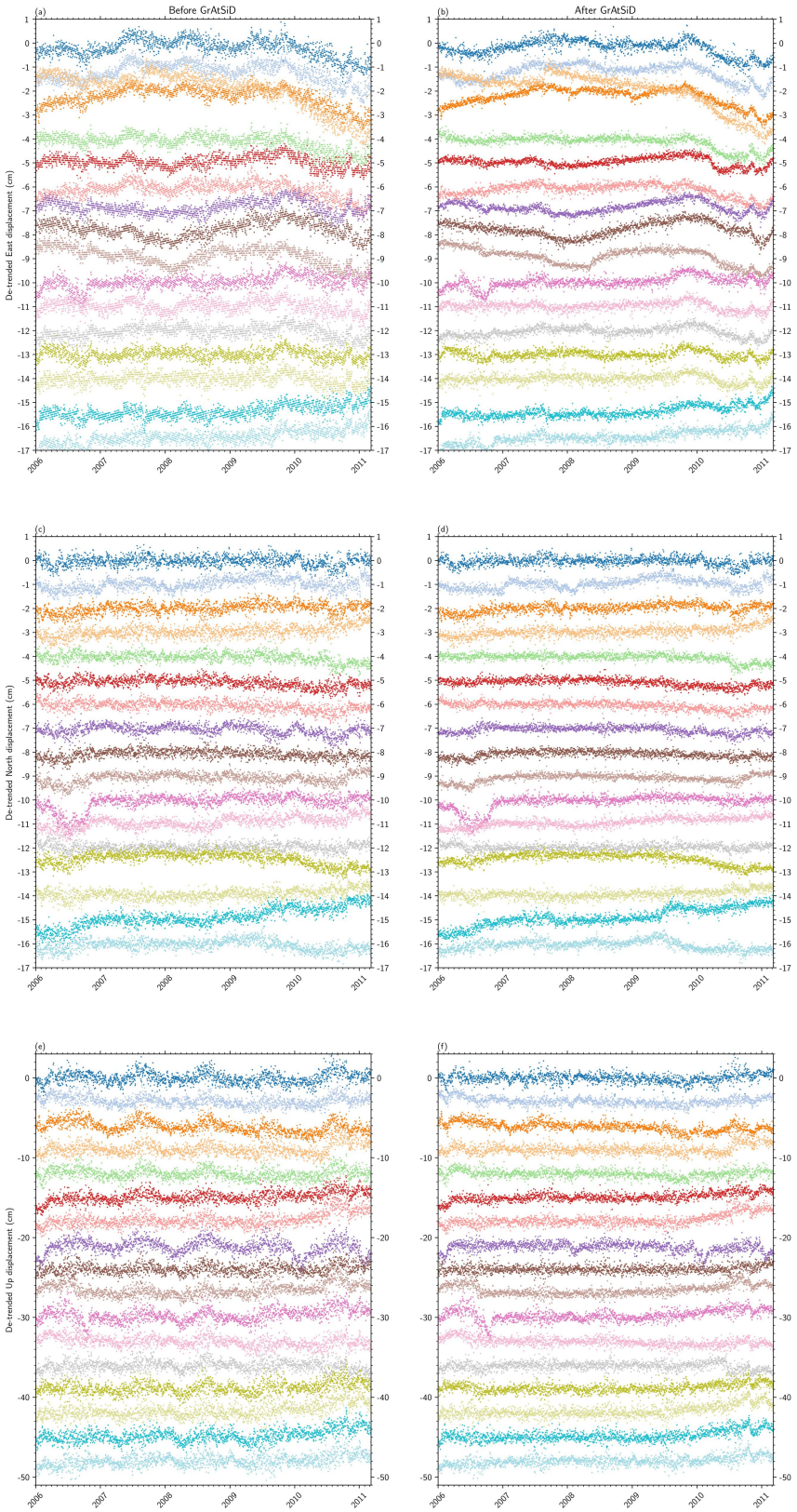
Author contributions Z.D. and J.C.B. processed the South American network solutions. J.R.B., M.M. and B.S. performed postprocessing (analysis of daily GNSS time series). M.B., J.C.B., Z.D. and J.R.B. investigated the processing artefacts and non-tectonic signals. T.J., O.O. and J.R.B. performed the geophysical and geological interpretation. J.R.B. did the kinematic modelling. All authors assisted in editing the manuscript.

Competing interests The authors declare no competing interests.

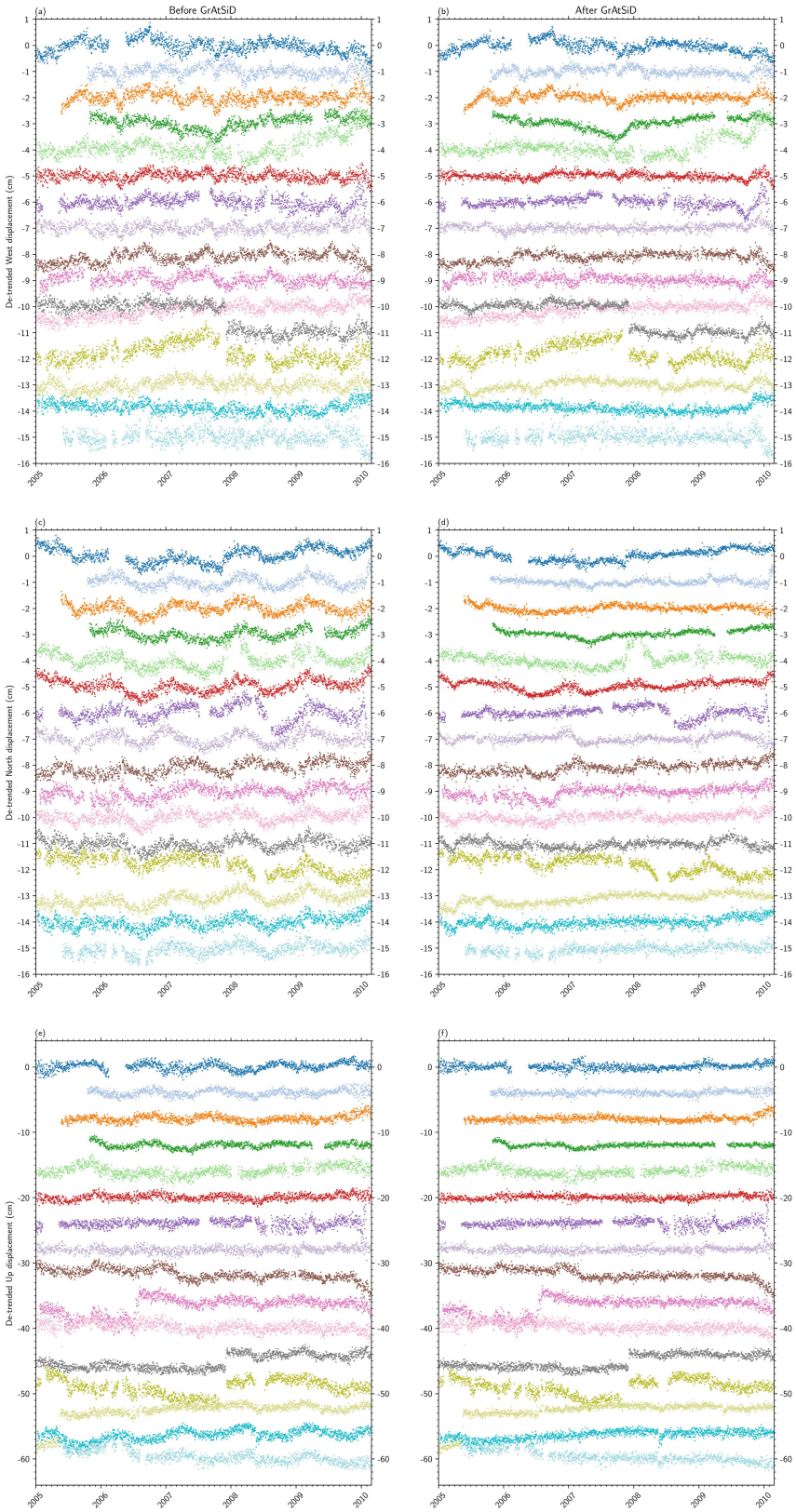
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2212-1.
Correspondence and requests for materials should be addressed to J.R.B.
Peer review information *Nature* thanks Yehuda Bock, Kosuke Heki and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



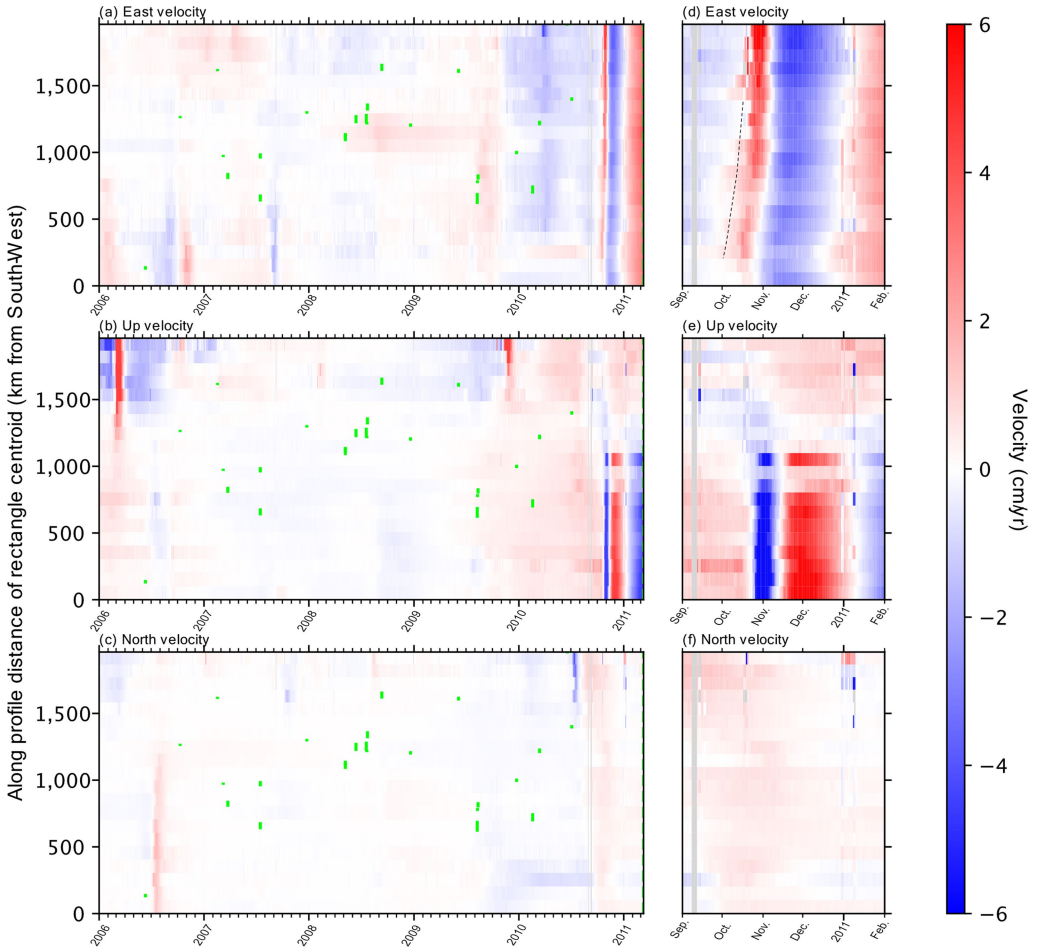
Extended Data Fig. 1 | Stations used in the processing and analysis of the pre-Maule-earthquake GNSS data. Locations of IGS stations used to define the reference frame of the network solutions are shown in gold. Black dots are the stations where network solutions are used in the time series analysis of the pre-Maule signals. Coloured triangles indicate locations of time series shown in Fig. 2 and Extended Data Fig. 3. There are some stations used to define the reference frame that are not used in the time series analysis owing to lack of data in the desired window (1 January 2005 until 25 February 2010).



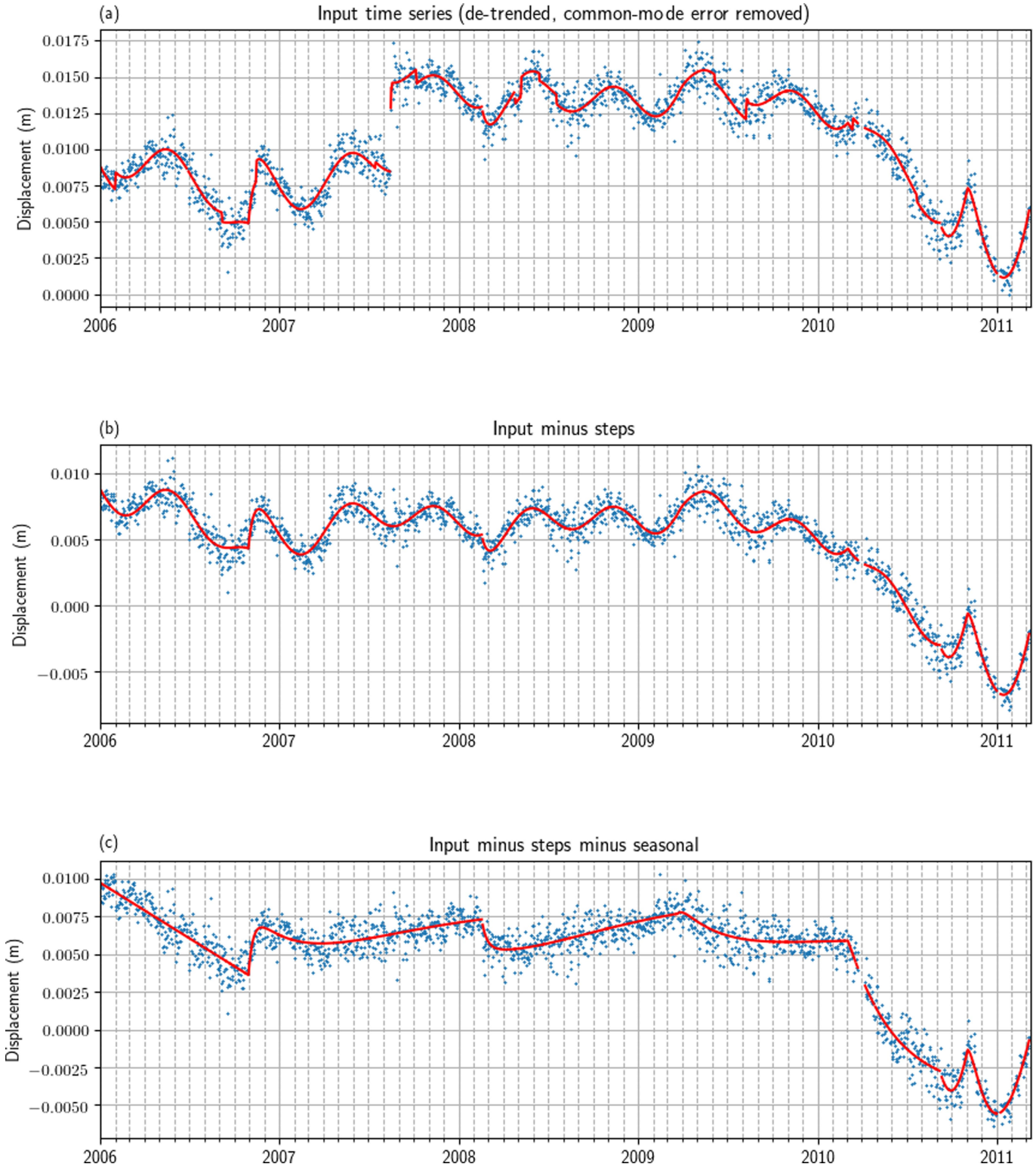
Extended Data Fig. 2 | Time series before the Tohoku-oki earthquake and the effect of noise removal with GrAtSiD in all three directional components. Left panels show the pre-Tohoku-oki F3 time series. Right panels show these time series after the removal of background seasonal and common-mode noise (with the GrAtSiD routine). The transient behaviour in the months before Tohoku-oki is heavily obscured by seasonal and common-mode noise. Colours correspond to locations on Fig. 1. For clarity, steps have been removed from all time series. Time series in these plots extend until three days before the mainshock.



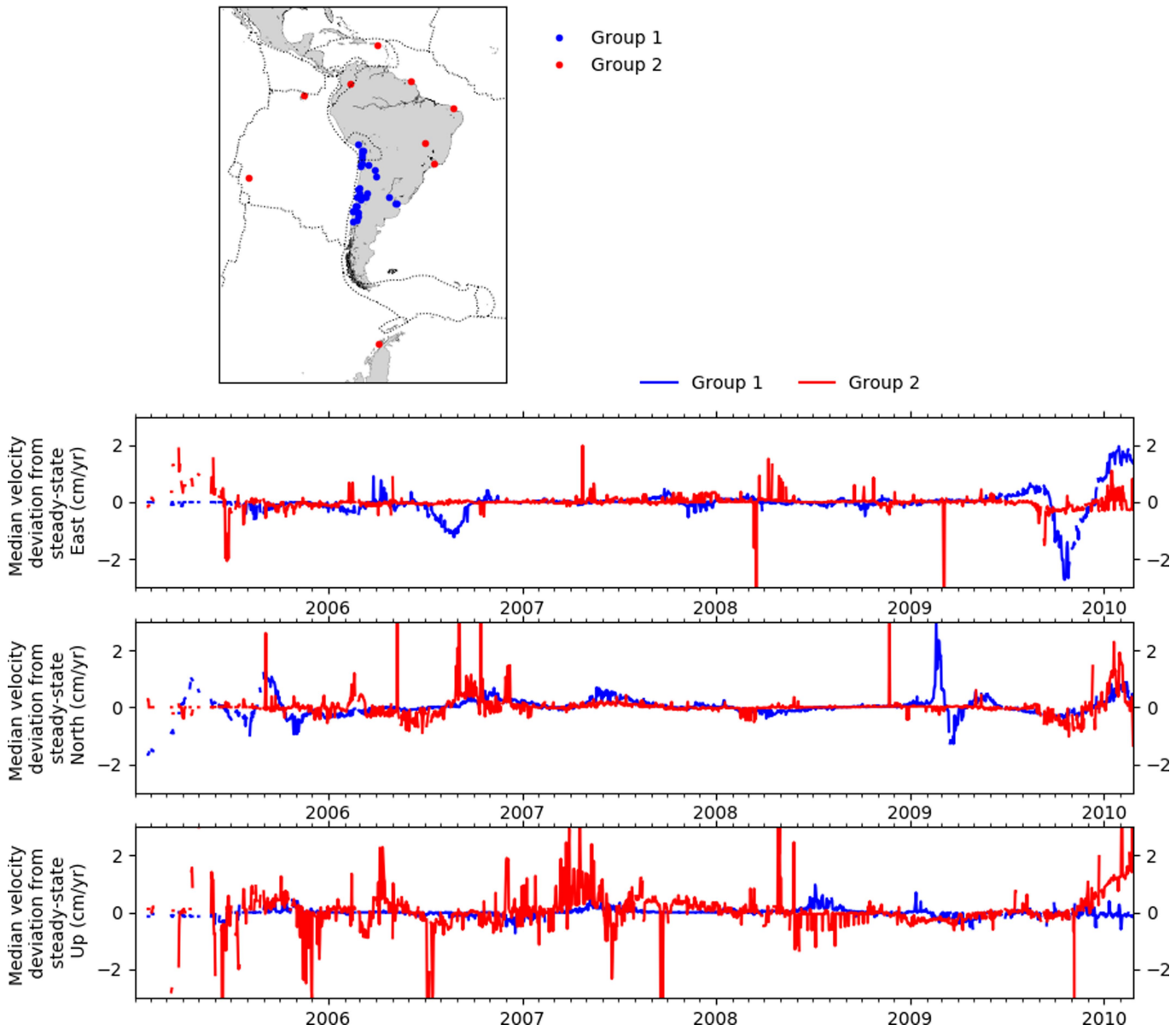
Extended Data Fig. 3 | Time series before the Maule earthquake and the effect of noise removal with GrAtSiD in all three directional components. Left panels show the pre-Maule time series. Right panels show these time series after the removal of background seasonal and common mode noise (with the GrAtSiD routine). The transient behaviour in the months before Maule event is heavily obscured by seasonal and common-mode noise. Colours correspond to locations on Fig. 1. For clarity, steps have been removed from all time series. Time series in these plots extend until two days before the mainshock.



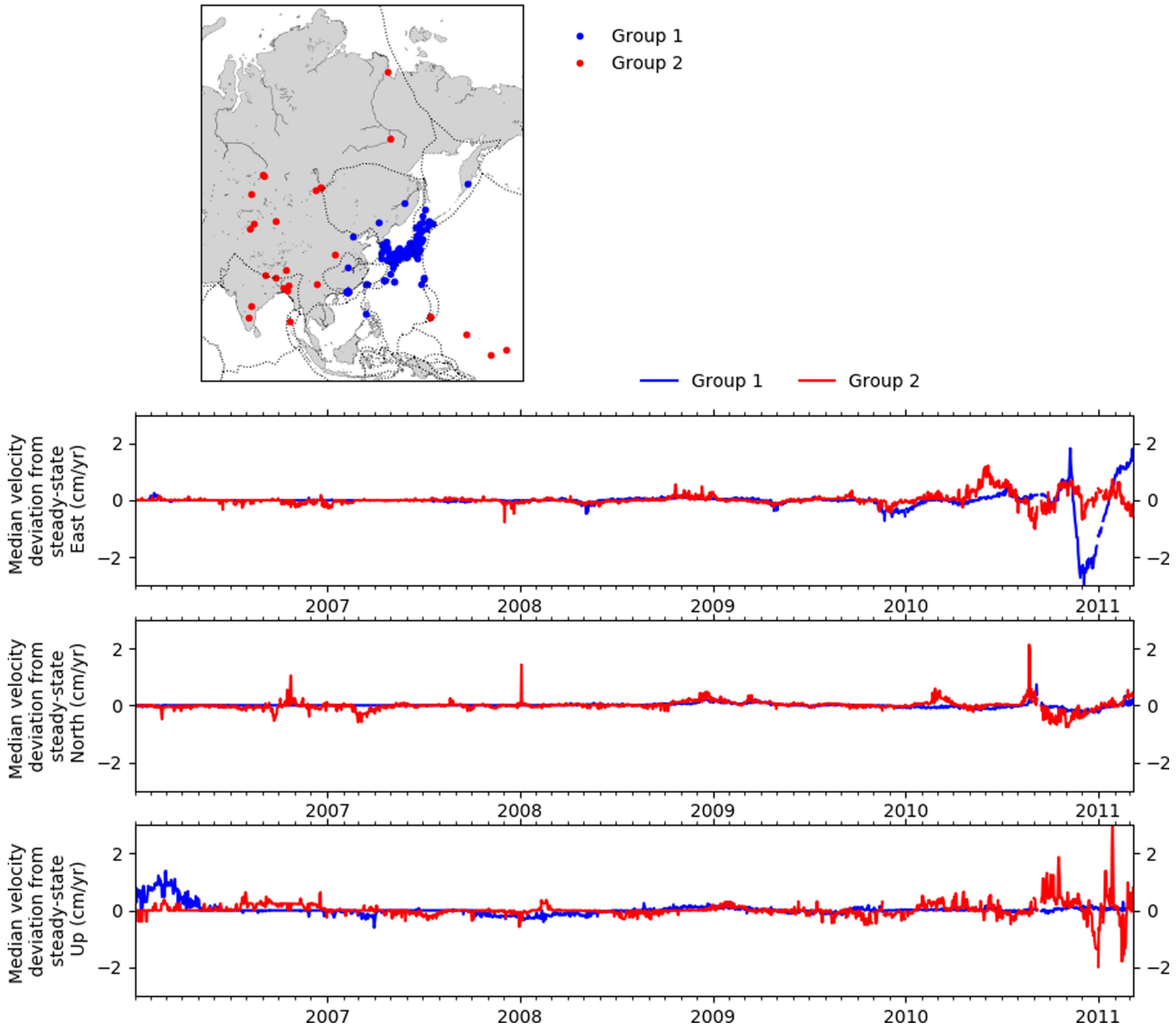
Extended Data Fig. 4 | Visualizing the along-strike signal migration and reversal of Japan in the years and months preceding the Tohoku-oki earthquake for all three directional components. Velocities within non-overlapping rectangular regions before the Tohoku-oki earthquake. The velocity for each region is detrended relative to the median velocity of that region between 1 January 2006 and 8 March 2011. Green lines indicate the along-strike locations and times of earthquakes of moment magnitude exceeding 6. Panels **d**–**f** are zoom-ins of panels **a**–**c** between the beginning of September 2010 and the beginning of February 2011. The dashed line on panel **d** indicates the velocity front that migrates across Japan from the southwest (shown in Supplementary Videos 3 and 4 and Fig. 3).



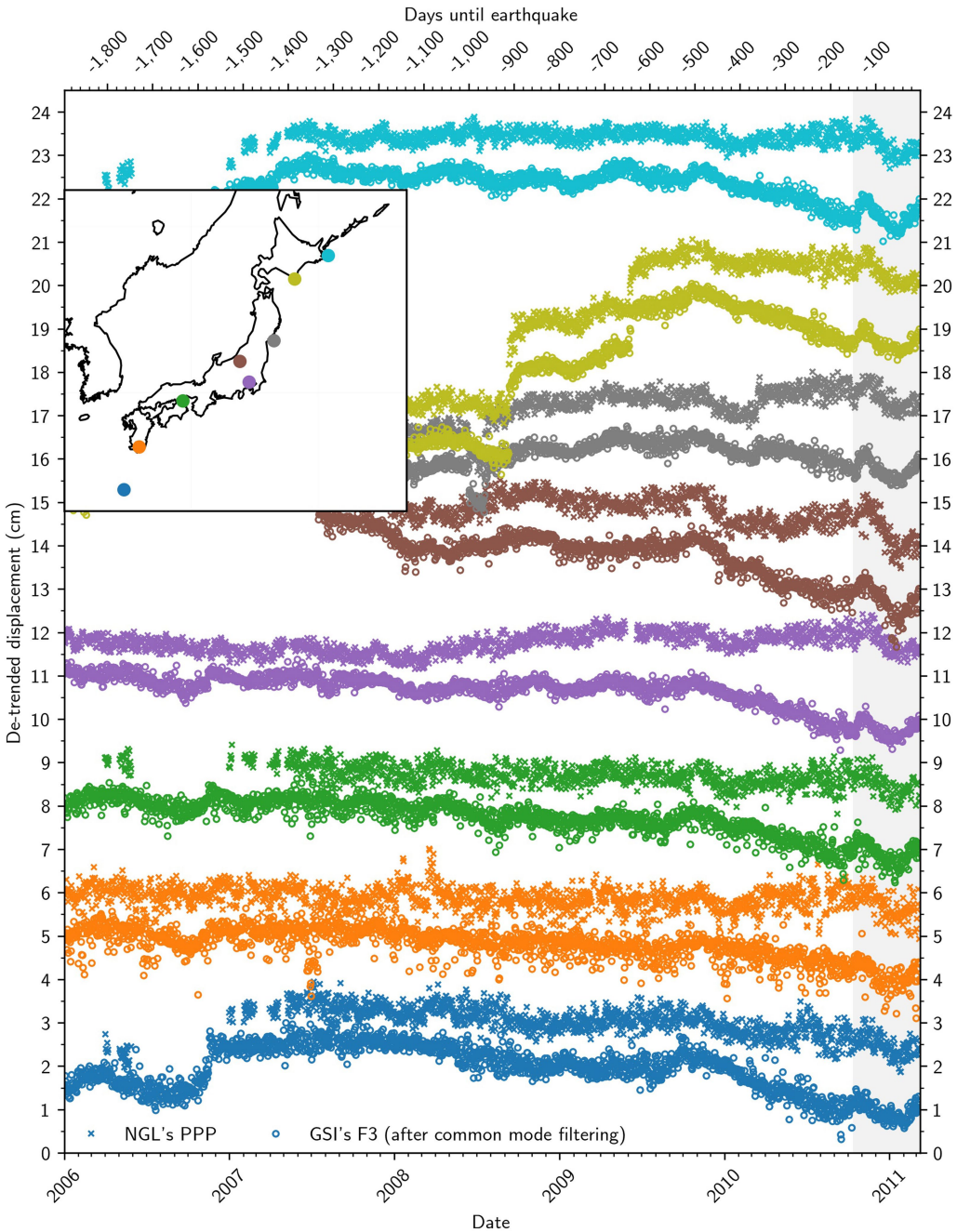
Extended Data Fig. 5 | The corrections to the time series made possible after application of the regression model solved by GrAtSiD. a, The example time series is for the East component of station Ooamishirasato in Japan. Blue dots show the time series input into the GrAtSiD routine. This time series has been corrected for the common-mode error. The red line shows the complete fit of the regression model solved by GrAtSiD. This includes steps, oscillation terms, the first-order polynomial and the multi-transients. The time series has been optimally tilted (detrended) for clarity of presentation. **b**, The time series (blue dots) and trajectory model (red line) after removal of the modelled step offsets. **c**, The time series and the regression model following the removal of the modelled seasonal and step terms. The remaining terms in the model are the first-order polynomial and multi-transients. It is these detrended modelled trajectories following seasonal and step removal (shown in panel **c** in red) that are represented in Supplementary Videos 3–8 and Figs. 3 and 4 and Extended Data Fig. 4.



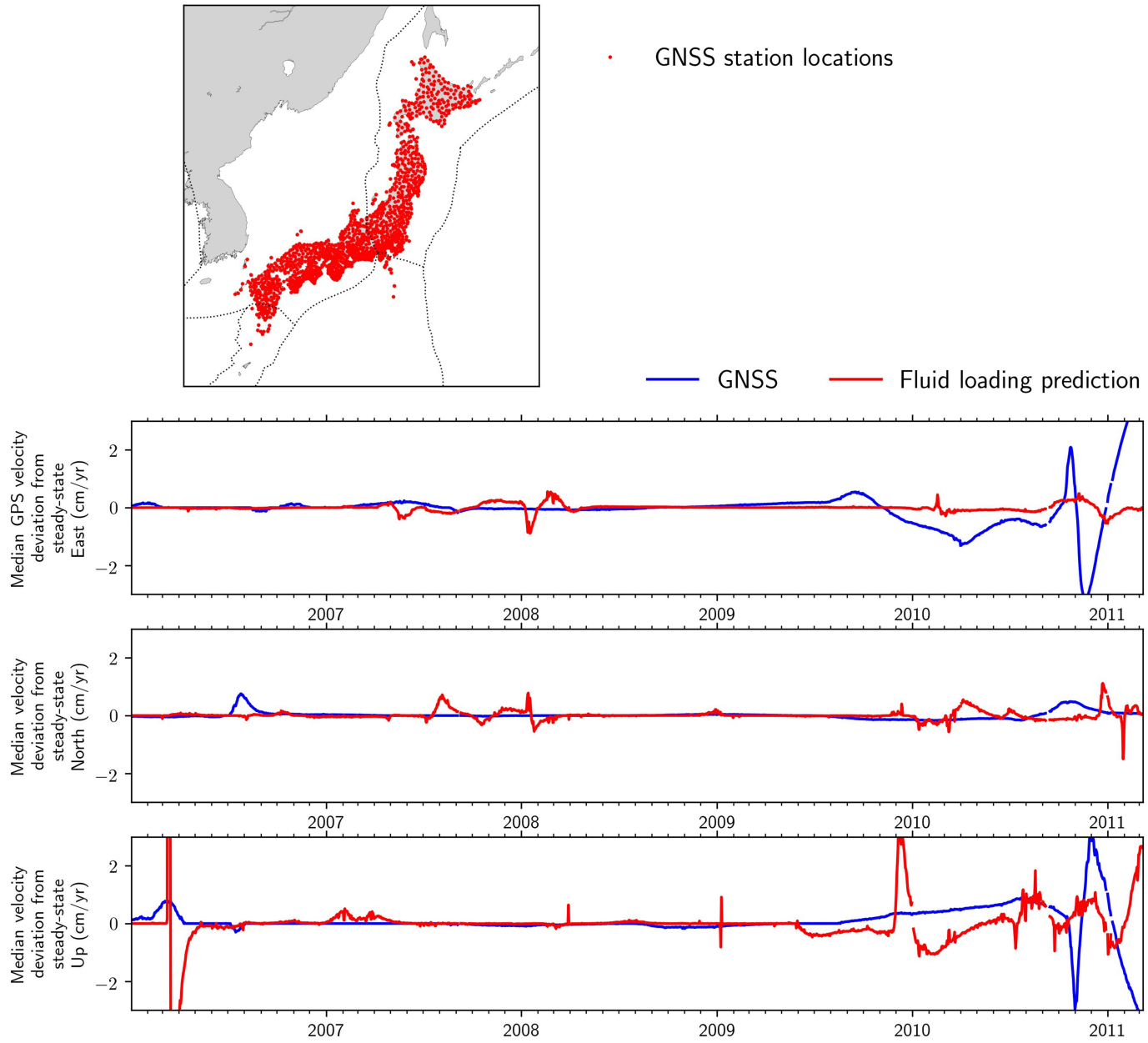
Extended Data Fig. 6 | Investigating spatial extents of the pre-Maule-earthquake wobbling in the network solutions. The map shows the locations of two groups of stations used in the investigation into spatial extent of the unstable period observed before the Maule earthquake. The time series show the average (median) deviation from median velocity at each station of the two groups in the above map for each directional component, where the median velocity of each station is determined between 1 January 2005 and 25 February 2010.



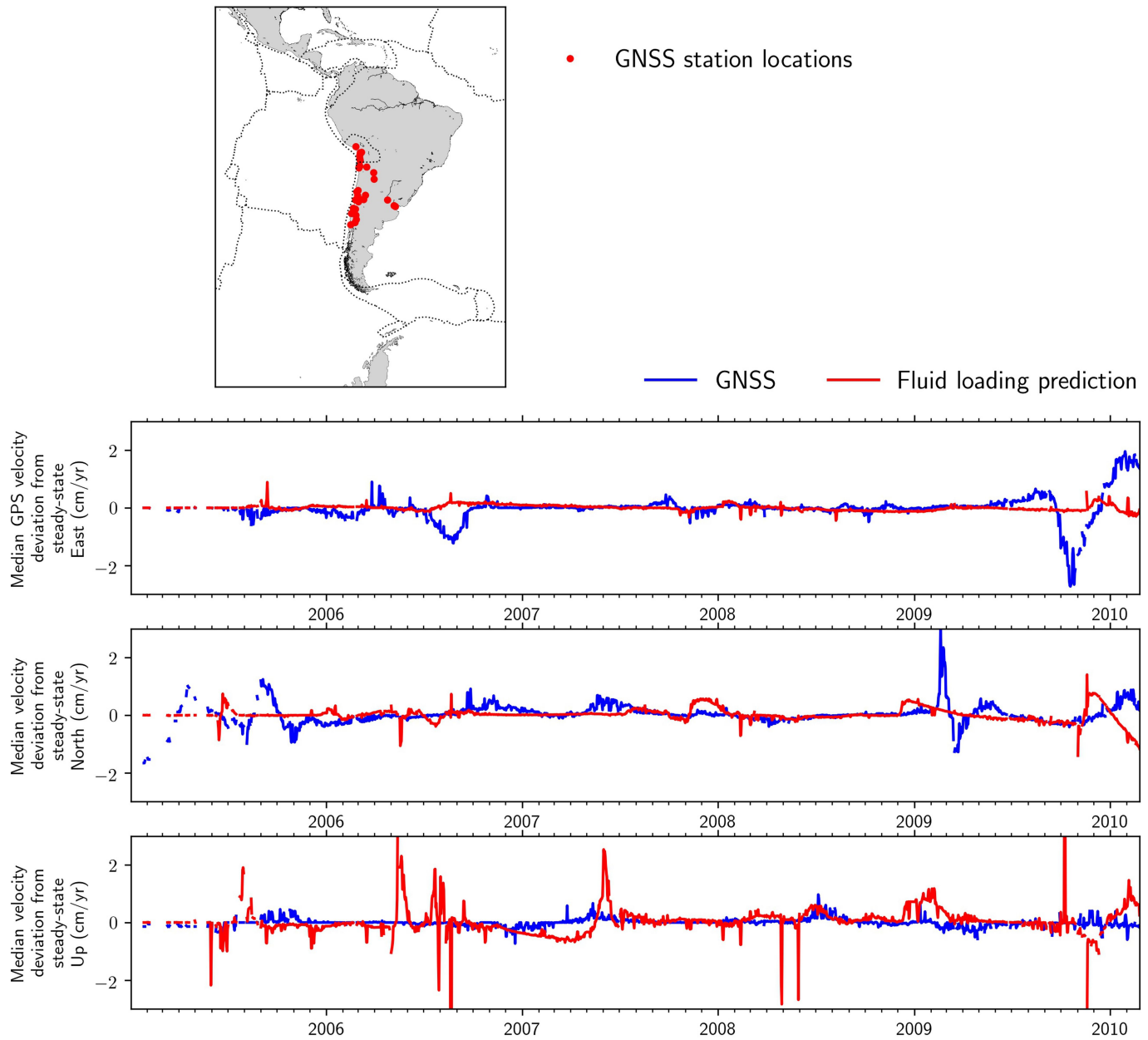
Extended Data Fig. 7 | Investigating spatial extents of the pre-Tohoku-oki-earthquake wobbling using the Nevada Geodetic Laboratory's IGS08 PPP solutions. The map shows the locations of two groups of stations used in the investigation into spatial extent of the unstable period observed before the Tohoku-oki earthquake. The time series show the average (median) deviation from median velocity at each station of the two groups in the above map for each directional component, where the median velocity of each station is determined between 1 January 2006 and 8 March 2011.



Extended Data Fig. 8 | Comparison of GSI's F3 solutions with NGL's IGS08 PPP solutions for selected stations across Japan before the Tohoku-oki earthquake. All time series shown are in the East component. Circles show the F3 and crosses show the PPP solutions. Colours correspond to the stations located on the inset map.



Extended Data Fig. 9 | Comparing the transient surface motions recorded by GNSS and predicted by fluid-loading models before the Tohoku-oki earthquake. The map shows the GNSS station locations used in the analysis comparing fluid-loading displacement predictions to GNSS displacement measurements for the pre-Tohoku-oki case. The time series show a comparison of the median velocity variations for GNSS-measured (GSI's F3 solutions) and fluid-loading-predicted displacements at the locations in the map. Velocities are taken from the trends estimated by GrAtSiD. In the horizontal components, the prediction from fluid loading produces much lower velocities than those observed. In the vertical component, there is considerable deviation from steady-state velocity in both the GNSS observation and fluid-loading prediction but with visibly low agreement in sense of motion.



Extended Data Fig. 10 | Comparing the transient surface motions recorded by GNSS and predicted by fluid-loading models before the Maule earthquake. The map shows the GNSS station locations used in the analysis comparing fluid-loading displacement predictions to GNSS displacement measurements for the pre-Maule case. The time series show a comparison of the median velocity variations for GNSS-measured and fluid-loading-predicted displacements at the locations in the map. Velocities are taken from the trends estimated by GrAtSiD. In the East component (in which the pre-Maule unstable motion is most pronounced) the prediction from fluid loading produces much lower deviation from steady-state velocities than those observed by GNSS.

Article

Measuring and forecasting progress towards the education-related SDG targets

https://doi.org/10.1038/s41586-020-2198-8

Received: 25 June 2019

Accepted: 18 March 2020

Published online: 15 April 2020

Open access

Check for updates

Joseph Friedman[1,2], Hunter York[1], Nicholas Graetz[1,3], Lauren Woyczynski[1], Joanna Whisnant[1], Simon I. Hay[1,4] & Emmanuela Gakidou[1,4]✉

Education is a key dimension of well-being and a crucial indicator of development[1–4]. The Sustainable Development Goals (SDGs) prioritize progress in education, with a new focus on inequality[5–7]. Here we model the within-country distribution of years of schooling, and use this model to explore educational inequality since 1970 and to forecast progress towards the education-related 2030 SDG targets. We show that although the world is largely on track to achieve near-universal primary education by 2030, substantial challenges remain in the completion rates for secondary and tertiary education. Globally, the gender gap in schooling had nearly closed by 2018 but gender disparities remained acute in parts of sub-Saharan Africa, and North Africa and the Middle East. It is predicted that, by 2030, females will have achieved significantly higher educational attainment than males in 18 countries. Inequality in education reached a peak globally in 2017 and is projected to decrease steadily up to 2030. The distributions and inequality metrics presented here represent a framework that can be used to track the progress of each country towards the SDG targets and the level of inequality over time. Reducing educational inequality is one way to promote a fairer distribution of human capital and the development of more equitable human societies.

The value of education is well-recognized, both as a primary human right and as a key driver of progress in economic development, health, fertility, politics, social empowerment, and human capital[3–13]. The international community recognized educational attainment as a key development priority in the Millennium Development Goals (MDGs), which became a key focus for a large variety of global actors. The education-related MDG targets focused largely on expanding primary education up to 2015[14], and great progress in this regard was seen as a result. In the SDGs—the follow-up to the MDGs with a target year of 2030—education was again highly prioritized, with a wider scope that emphasized reducing inequalities.

Increases in global schooling rates

SDG target 4.1 calls for universal primary schooling. Progress towards this goal has been, and is projected to continue to be, substantial (Fig. 1). Globally, the proportion of 25–29-year olds with at least 6 years of schooling rose from 50.1% (95% uncertainty interval: 49.3–51.0%) in 1970 to 83.2% (82.1–84.0%) in 2018 and is projected to reach 89.4% (87.4–91.0%) by 2030. Even as far back as 1970, countries in high-income regions and in eastern Europe and central Asia had on average already achieved near universal primary attainment. In the remaining regions, rates of primary attainment have risen substantially. Although this progress is to be celebrated, important gaps remain in a subset of nations that are not projected to achieve near universal levels of primary attainment by 2030, largely due to gaps in schooling among women (see Extended Data Fig. 1).

SDG target 4.1 also calls for universal secondary schooling. However, secondary attainment estimates reveal a much more heterogeneous picture. In 1970, countries generally fell into one of two categories; nearly 50% of the global population aged 25–29 residing in highly educated regions had already attained 12 years of schooling, whereas the rest of the world saw rates at or below 10%. Although global attainment of at least 12 years of schooling has risen steadily since 1970, no major world region has achieved near universal levels. All regions have seen progress, yet the inter-regional disparities remain massive in 2018 and are projected to decrease only slightly in the coming years.

SDG target 4.3 addresses tertiary education, calling for 'equal access' for all individuals. Tertiary education exhibited a substantial scale-up between 1970 and 2018 that is projected to continue in the coming decade, although global completion rates remain low. Similar to the trend in secondary education, the high-income and eastern European and central Asian regions exhibit substantially higher rates throughout the time period shown, and are projected to achieve about half of their population completing tertiary education by 2030. The remaining regions have also seen progress, with much of the growth seen after 2000. The increase is particularly notable in North Africa and the Middle East as well as in Southeast Asia, East Asia, and Oceania.

In summary, regional disparities in tertiary education completion are increasing over time and are projected to continue to do so, whereas secondary gaps are expected to decrease only slightly. The success of narrowing the global gap for primary education has not been extended to higher levels of education, which raises concerns

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA. [2]Center for Social Medicine and Humanities, University of California Los Angeles, Los Angeles, CA, USA. [3]Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Health Metrics Sciences, School of Medicine, University of Washington, Seattle, WA, USA. ✉e-mail: gakidou@uw.edu

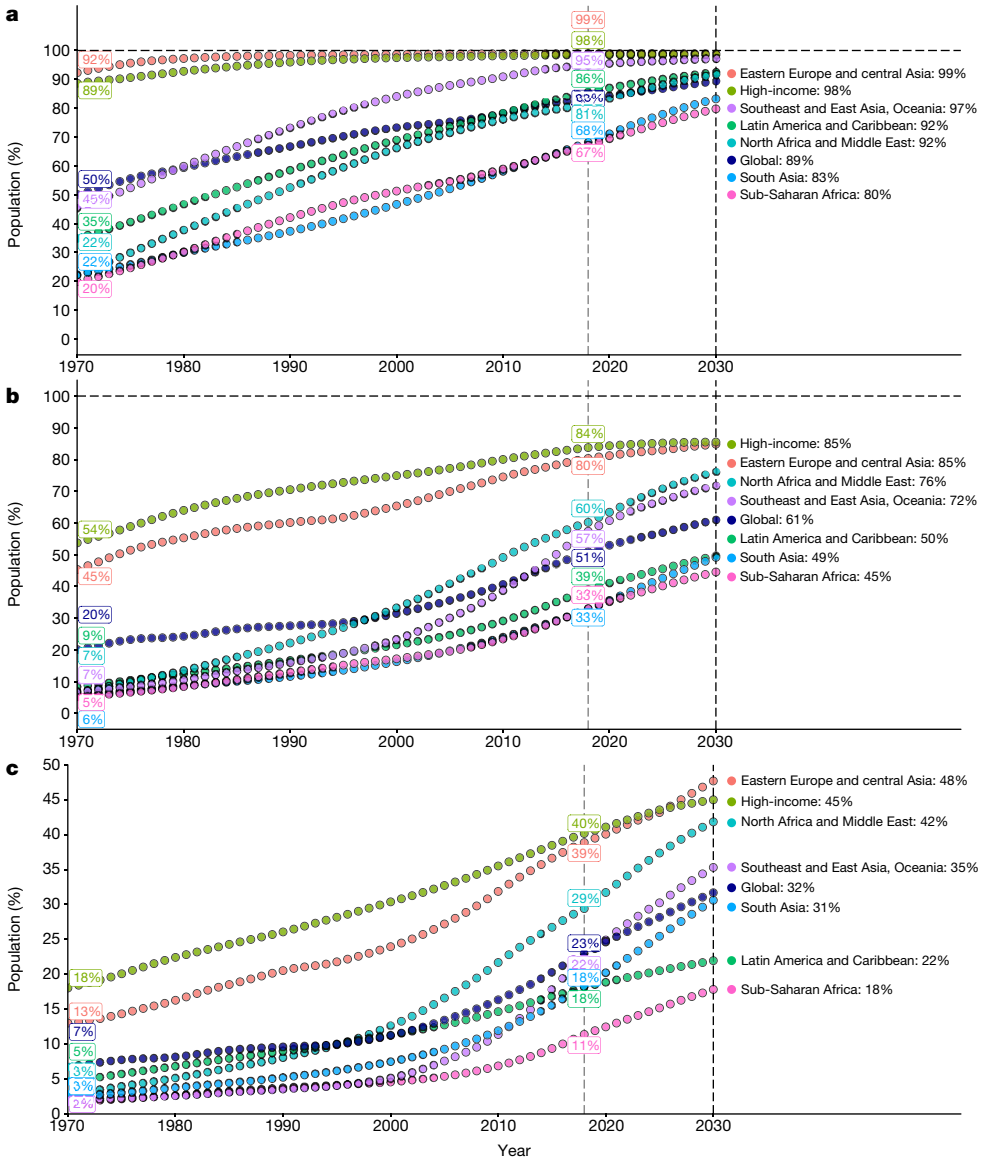


Fig. 1 | Regional attainment of primary, secondary, and tertiary schooling from 1970 to 2030. a–**c**, Attainment rates of 6+ (**a**), 12+ (**b**), and 15+ (**c**) years of schooling are shown. All trends reflect 25–29-year-old individuals separated by major world region. The vertical dashed lines indicate 2018, when the forecasts begin, and 2030, the target year for the SDGs.

about gaps in opportunities amplifying across regions in the coming decade.

Progress towards gender equity

Gender equity has been a central focus of the SDG targets. SDG target 5 calls for gender equity broadly, and target 4.5 calls for the elimination of all gender disparities in education. We find that great strides have been made in reversing educational disparities for women globally, and in all regions of the world.

To benchmark the progress of each country towards gender parity in education, we calculate the absolute gap in the mean years of schooling, and assess the contribution of primary, secondary, and tertiary schooling to these gaps (Fig. 2). In 1970, men aged 25–29 years had completed on average 1.7 (1.6–1.8) additional years of education compared with women of the same age. By 2018, this gap had nearly closed, falling to only 0.3 (−0.2–0.8) years, and is projected to reverse by 2030. Previous modelling studies of global gender differences in educational attainment that have focused on all adults 25 and older show progress, but note that women are not yet close to catching up to men[15]. By focusing only on young women and men, we show that among the most recently educated members of societies, women had in fact nearly closed the gender gap in 2018. Young men had statistically significantly higher levels of attainment compared with women, at the 95% confidence level, in 142 countries in 1970, 27 countries in 2018, and only 4 countries by 2030. For 2030, the countries in which women's education is predicted to still lag behind that of men are predominantly in sub-Saharan Africa, Southeast Asia, East Asia, and Oceania. In addition, by 2030, women are expected to achieve statistically significantly higher mean years of schooling than men in 18 countries—a tremendous reversal of the global landscape that was observed in 1970.

In absolute terms, the largest component of this reduction has been observed in primary education. In 1970, men aged 25–29 completed 0.9 (0.9–1.0) additional years of primary schooling compared with women, which fell to only 0.3 (0.2–0.4) years in 2018. This reflects progress in nearly every region; all had primary education gaps favouring men in 1970. By 2018, these gaps had shrunk by considerable margins in every region, and many disappeared entirely. Nevertheless, a small number

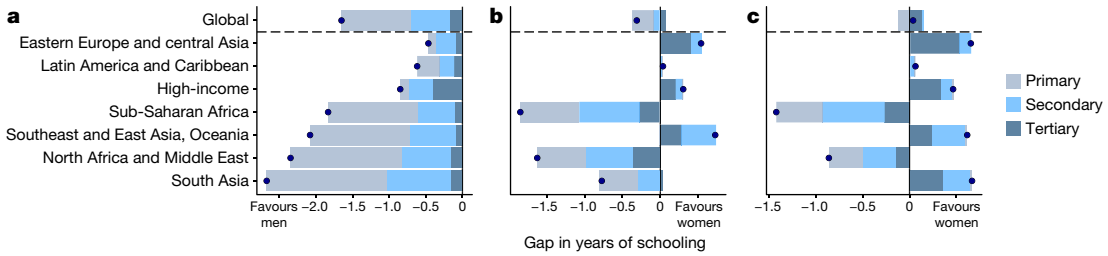


Fig. 2 | Regional gender gaps in primary, secondary, tertiary, and total schooling. a–c, The gender gap is shown for 1970 (**a**), 2018 (**b**), and 2030 (**c**). The total gap in years of schooling is represented by a dot, for individuals aged 25–29, separated by each regional group. The grey, light blue, and dark blue bars represent the contributions of primary, secondary, and tertiary schooling, respectively, to the total gender gap.

of countries are forecast to have persistent gaps in attainment of at least 6 years of schooling, largely in North Africa and the Middle East, as well as sub-Saharan Africa (Supplementary Fig. 14).

Secondary and tertiary education both show a more heterogeneous pattern, in which women are overtaking men in most regions of the world, whereas large-magnitude disparities seen in sub-Saharan Africa and North Africa and the Middle East are projected to persist. Our estimates indicate that in 2012, women aged 25–29 overtook men in the global average of tertiary attainment, and they are forecast to do so for secondary attainment in 2026. Unlike primary attainment, which has largely converged globally in a place of gender parity, women have overtaken men by substantial margins in many nations in Latin America, Asia, and Europe. This phenomenon has been reported for many nations in the Organisation for Economic Co-operation and Development (OECD)[16,17] and elsewhere[18–20], in which boys increasingly fall behind girls in schooling as nations develop. Our results indicate the commonality of this trend for many regions of the world, and show how these advances have contributed to closing the overall gender gap. Notably, our results indicate that these gaps are projected to grow with time.

Assessment of inequalities in education

Although gender equity is of crucial importance, it only captures one dimension of inequality in education. Beyond gender, SDG target 4.5 calls for broad social equity in educational attainment, across lines of ethnicity, race, socio-economic status, ability, and other identities[21]. The particular social groupings that are relevant vary across countries, but insight between countries can be gleaned by assessing the total inequality.

To facilitate benchmarking between nations and a global assessment of trends in educational inequality, we use a metric of the total within-country inequality in education, the average interpersonal difference (AID), which represents the average difference between any two individuals in a population. Results and discussion using alternative metrics of inequality, including relative measures such as the Gini coefficient, are presented in the Supplementary Information.

Globally, inequality rose steadily before peaking in 2017 with a 4.6-year (4.5–4.7) average within-country difference between any two given individuals (Fig. 3a). Subsequently, inequality has been decreasing and is projected to continue to do so up to 2030. Looking at the arc of inequality in education over time across regions and countries, a consistent Kuznets curve can be observed in almost every setting. A Kuznets curve describes a development trend in which progress is associated with first increased and then decreased inequality, creating an inverse-U-shaped curve[22].

We observe substantial variation in the maximum level of inequality reached during each period, which in some cases reflect threefold differences in the degree of equality for a given average level of schooling. In this way, these curves provide a valuable tool for comparing the level of inequality of each country compared with their neighbours, relative to their overall level of progress.

Latin America and the Caribbean had the highest levels of inequality in 1970, with an AID of 4.5 years (4.4–4.6) (Fig. 3a). Over time, however, Latin America and the Caribbean has had an only intermediate-height Kuznets curve, despite substantial progress in the mean years of educational attainment (Fig. 3b). Latin America and the Caribbbean stands out as having less inequality in education at each point in the development arc compared with regions such as South Asia or North Africa and the Middle East, as shown by a lower overall Kuznets curve. This result highlights the need to assess inequality for each region with respect to its level of development by looking across decades to understand variation in the arc of educational expansion.

Between 1970 and 2018, sub-Saharan Africa and South Asia saw great advances in education, and also large increases in inequality. South Asia had the highest level of educational inequality globally in 2018, with an AID value of 6.0 (5.7–6.3). Its Kuznets curve is largely similar to that of North Africa and the Middle East. If sub-Saharan Africa continues to develop at its current trajectory, we expect its trend to look similar to that of Latin America and the Caribbean, which is approximately 30 years further along the development arc. Taking a more granular look, substantial variation can be seen between nations in sub-Saharan Africa (Fig. 3c). In 2018, several countries in western sub-Saharan Africa displayed the highest inequality values in the world, well above the ninetieth percentile mark for their level of mean attainment. Nevertheless, several nations in southern Africa are below the tenth percentile of inequality values for their mean attainment.

The region of Southeast and East Asia and Oceania is noteworthy for having the flattest Kuznets curve, and therefore the least unequal trajectory of development among low- and middle-income countries. Eastern Europe and central Asia underwent rapid gains in education from 1970 to 1995, achieving mean values similar to high-income countries by 2018, with lower overall inequality.

Centring equality in global progress

Educational inequalities exist in many different forms and need to be addressed in order for societies to maximize well-being and the potential for education to facilitate economic development. Gender gaps are projected to persist for girls in much of the developing world and widen for boys in a subset of developed countries[16,23]. Disparities can also be found along dimensions of wealth, ethnicity, race, ability, and other social groupings[20,24]. Previous work has shown substantial inequalities in education between urban and rural areas[5,25,26], and along lines of wealth[20]. These inequalities are easy to miss when drawing on national average measures of attainment.

The distributions and inequality metrics presented here provide a framework that can be used to track the progress of each country towards the SDG targets and levels of inequality over time. Once detected, inequalities can be reduced with the implementation of specific policies. For example, eliminating school fees, improving local access to schools, increasing the number of years of compulsory schooling, and providing food, stipends, and other resources for children at school are known to increase participation among the most economically disadvantaged children, and the creation of special governmental bodies can reduce gaps for children of minority ethnic groups[25,27,28]. It is therefore essential to examine progress in average levels of attainment with an understanding

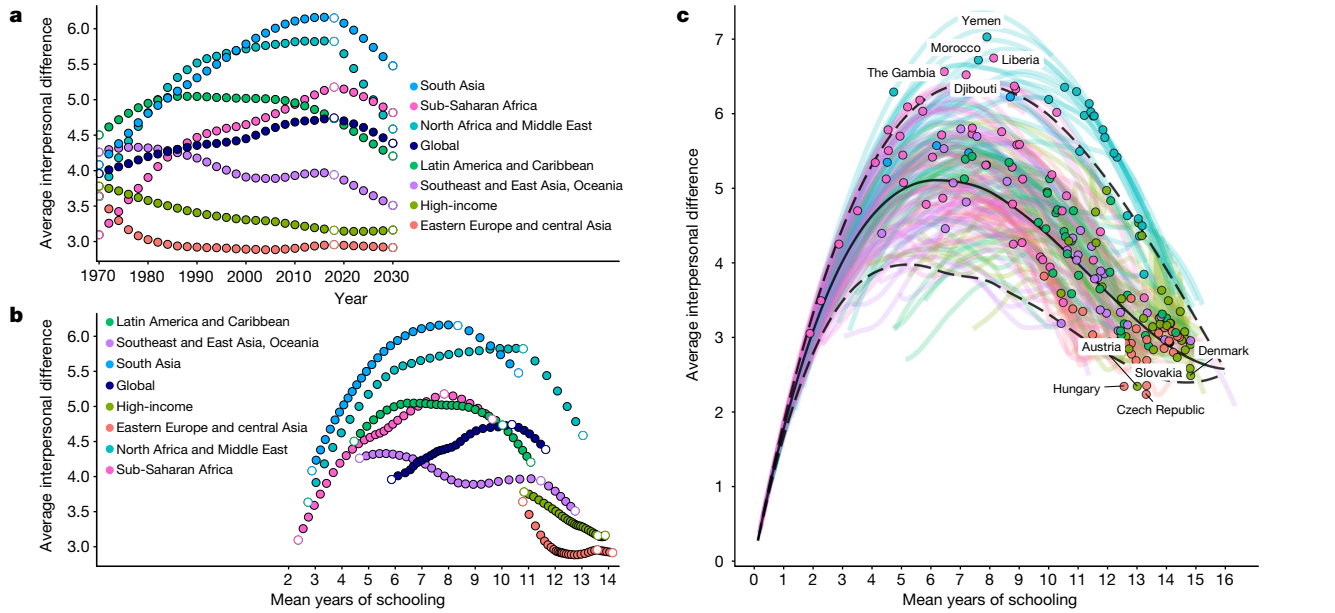


Fig. 3 | Trajectories in educational inequality. a, Trends in educational inequality are shown over time, with labels indicating the rank of inequality levels in 2030. **b**, Trends in educational inequality are shown with respect to mean years of schooling, with labels indicating the rank of inequality levels in 1970. Results are shown globally and regionally for every second year from 1970 to 2030, for individuals aged 25–29. The white dots mark 1970, the beginning of the estimates, 2018, the beginning of forecasts, and 2030, the SDG target year. **c**, National trends in the AID and mean years of schooling are shown from 1970 to 2030, with the value for 2018 shown as a bold point. The five highest and lowest values in 2018 are labelled. The solid line shows the median level of inequality for a given degree of mean years of schooling, across all years of data from 1970 to 2030, and the dashed lines show the smoothed ninety-fifth and fifth percentiles. Quantiles were calculated over modelled estimates from $n = 195$ countries.

of the full within-country distribution and inequality. Gains in education are linked to improvements in numerous other sectors of society[3,4,13]. Ensuring equality in education will translate into positive effects in the equality of human productivity, health, and well-being.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2198-8.

1. UNESCO. *Migration, Displacement and Education: Building Bridges, not Walls*. https://en.unesco.org/gem-report/report/2019/migration (2018).
2. Marmot, M., Friel, S., Bell, R., Houweling, T. A. & Taylor, S. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* **372**, 1661–1669 (2008).
3. Lim, S. S. et al. Measuring human capital: a systematic analysis of 195 countries and territories, 1990–2016. *Lancet* **392**, 1217–1234 (2018).
4. Gakidou, E., Cowling, K., Lozano, R. & Murray, C. J. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. *Lancet* **376**, 959–974 (2010).
5. Graetz, N. et al. Mapping local variation in educational attainment across Africa. *Nature* **555**, 48–53 (2018).
6. UNESCO Institute for Statistics. *Quick Guide to Education Indicators for SDG 4*. http://uis.unesco.org/sites/default/files/documents/quick-guide-education-indicators-sdg4-2018-en.pdf (2018).
7. Nilsson, M., Griggs, D. & Visbeck, M. Policy: map the interactions between sustainable development goals. *Nature* **534**, 320–322 (2016).
8. Lutz, W., Cuaresma, J. C. & Sanderson, W. Economics. The demography of educational attainment and economic growth. *Science* **319**, 1047–1048 (2008).
9. Basu, A. M. & Stephenson, R. Low levels of maternal education and the proximate determinants of childhood mortality: a little learning is not a dangerous thing. *Soc. Sci. Med.* **60**, 2011–2023 (2005).
10. Bicego, G. T. & Boerma, J. T. Maternal education and child survival: a comparative study of survey data from 17 countries. *Soc. Sci. Med.* **36**, 1207–1227 (1993).
11. Hatt, L. E. & Waters, H. R. Determinants of child morbidity in Latin America: a pooled analysis of interactions between parental education and economic status. *Soc. Sci. Med.* **62**, 375–386 (2006).
12. Lutz, W. & Kc, S. Global human capital: integrating education and population. *Science* **333**, 587–592 (2011).
13. The European Commission. *Demographic and Human Capital Scenarios for the 21st Century* (eds Lutz, W. et al.) (2018).
14. United Nations. *The Millenium Development Goals Report 2015*. http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf (2015).
15. Goujon, A. et al. A harmonized dataset on global educational attainment between 1970 and 2060 – an analytical window into recent trends and future prospects in human capital development. *J. Demogr. Economics* **82**, 315–363 (2016).
16. OECD. *PISA 2015 Results (Volume I)* (2016).
17. OECD. *The Pursuit of Gender Equality* (2017).
18. Lopus, S. & Frye, M. Visualizing Africa's educational gender gap. *Socius* **4**, 237802311879595 (2018).
19. Grant, M. J. & Behrman, J. R. Gender gaps in educational attainment in less developed countries. *Popul. Dev. Rev.* **36**, 71–89 (2010).
20. Jones, G. W. & Ramchand, D. S. Closing the gender and socio-economic gaps in educational attainment: a need to refocus: closing the gender and socio-economic gaps in educational attainment. *J. Int. Dev.* **28**, 953–973 (2016).
21. United Nations. *#Envision2030 Goal 4: Quality Education*. https://www.un.org/development/desa/disabilities/envision2030-goal4.html (accessed November 2018)
22. Kuznets, S. Economic growth and income inequality. *Am. Econ. Rev.* **45**, 1–28 (1955).
23. UNESCO. *Achieving Gender Equality in Education: Don't Forget the Boys*. https://unesdoc.unesco.org/ark:/48223/pf0000262714 (2018).
24. Lewis, M. & Lockheed, M. *Inexcusable Absence: Why 60 Million Girls Still Aren't in School and What to do About It* (Center for Global Development, 2006).
25. UNICEF. *Magic Box - School Mapping*. http://school-mapping.azurewebsites.net/ (accessed November 2018).
26. Graetz, N., Woyczynski, L., Wilson, K. F., Hay, S. I. & Gakidou, E. Mapping persistent local disparity in educational attainment. *Nature* **555**, 48–53 (2019).
27. Urbina, D. R. Intergenerational educational mobility during expansion reform: evidence from Mexico. *Popul. Res. Policy Rev.* **37**, 367–417 (2018).
28. Cohen, A. K. & Syme, S. L. Education: a missed opportunity for public health intervention. *Am. J. Public Health* **103**, 997–1001 (2013).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Overview

Our study follows the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER)[29]. We use a multi-stage model to estimate the average years of schooling, and the single-year distribution of educational attainment, for 1970 to 2018, and create projections to 2030. These models draw on a database of 3,180 nationally representative censuses and surveys. Estimates are created for the 195 counties and territories examined in the Global Burden of Disease 2017 study[30]. In the first stage, we model mean years of schooling and the proportion of the population without any formal schooling from 1970 to 2018. This is performed using a cohort extrapolation model and a subsequent age period model with Gaussian process regression to synthesis all data and create final estimates with uncertainty. The second stage entails an ensemble *K*-nearest neighbours algorithm to estimate the distribution of education from 1970 to 2018, drawing on previously estimated quantities. Finally, trends in these distributions are projected to 2030 using a rate of change approach, and mean years of schooling values for 2019–2030 are calculated from the resulting distributions. All analyses are run using 1,000 draws to propagate model and data uncertainty through to subsequent steps. All estimation steps are validated, and all hyper-parameters are optimized, using out of sample predictive validity.

Data sources

We compiled a database of 3,180 nationally representative surveys and censuses describing the distribution of years of schooling by age and sex. Data sources providing single years of schooling are used directly, while those providing larger bins of educational attainment, for example 'some primary attainment' are probabilistically split into single-year proportions using a previously published crosswalk model[31]. Data are top-coded to 18 years, as it is a common choice among providers of single-year education data[32], and it is reasonable to assume that the importance of education for health or social capital diminishes greatly after the completion of 18 years, which represents 2 to 3 years of post-university education in most educational systems.

Data adjustment model

Data are adjusted for systematic biases between data providers in a regional and location-specific fashion. Gold-standard data are identified using expert knowledge of the high-volume data providers that have robust processes in place to ensure data quality. In almost all cases, census data obtained from the IPUMS data repository are considered as the gold standard, or Demographic Health Survey data where IPUMS are not available. Supplementary Table 3 lists the location-specific gold-standard data providers. Regional effects are applied to all data to adjust them to the gold standard available in that region. Subsequently, in countries that had gold-standard data available, country-specific effects are used to adjust for within-county biases between data sources. This has the benefit of being able to correct for biases in all countries, even when gold-standard data are not available in that country, using regional effects. Country-specific effects ensure consistent time trends with minimal discontinuities.

We use a mixed-effects regression model with random effects for data provider and nested random effects for data provider within country. This model is run separately for each region, and is formulated as follows:

$$\text{logit}(P_{Q,A,S,Y,L}) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{sex} + \beta_3 \times \text{location} + \beta_4 \times \text{year} + u_{\text{dataprovider}} + u_{\text{location:dataprovider}}$$

in which $P_{Q,A,S,Y,L}$ is the quantity of interest, either proportion of the population with no education or $\frac{\text{mean years of educational attainment}}{18}$ for a given age, sex, year, and location. $u_{\text{data provider}}$ is a region-specific random effect that captures the average bias between data providers across all countries within that region, and $u_{\text{location:data provider}}$ is a nested location-specific random effect that captures the additional bias between a location-specific gold standard (where applicable) and the other sources present in that location.

To calculate source adjustments for each data provider, the $u_{\text{data provider}}$ value for each data provider is compared with the regional gold standard, and the difference is applied to all surveys. Subsequently, in locations that have gold-standard data present, $u_{\text{location:data provider}}$ effects are applied in the same fashion.

Cohort extrapolation

We use an age-cohort modelling process to project cohorts through time, leveraging the stability of cohort-specific educational attainment after age 25. To model the changes by age within cohorts, we use data from all available cohorts with multiple observations at or after age 25. For each quantity being modelled, we calculated y_{Q,L,S,C,A_x}, which is the logit difference of the $P_{Q,A,S,Y,L}{}'$ (the adjusted input data) at time x and at time y, for all possible combinations of repeat cohort observations. We restrict repeat cohort observations to those that are less than or equal to 10 years apart and to those where both observations occur after 1990 to avoid the attribution of differences in measurements to mortality as opposed to advances in survey and census design. In addition, we normalize all repeat cohort observation pairings so that the average change at 65 years of age is 0 to account for systematic bias between survey iterations (such as improvements in sampling). This is similar to other previously described approaches[33], in which only excess mortality beyond the age of 65 is considered. This calculation is shown below:

$$y_{Q,L,S,C,A_x} = \text{logit}(P_{L,S,C,A,\text{Src}_x}{}') - \text{logit}(P_{L,S,C,A,\text{Src}_y}{}') - \text{bias}_{L,P,S,C,\text{Src}}$$

in which Q is the quantity being modelled, L is location, S is sex, C is cohort, A is age, Src is data provider, and $\text{bias}_{L,P,S,C,\text{Src}}$ is the average change for cohorts as they age from 60 to 70 between the two surveys. This is the age period for which we expect the educational attainment of a cohort to be least prone to changes due to migration and mortality, and any changes observed during this period are therefore used as a measure of inherent bias between multiple waves of a survey or census.

These logit differences were examined with respect to several predictor variables. We then modelled the logit difference using a number of linear mixed-effects models, which were evaluated using out-of-sample predictive validity (see Supplementary Information). The best performing model specification is displayed here:

$$y_{Q,L,S,C,A_x} = I + u_{\text{location :super region}}$$

in which I is a natural spline with a knot at age 70 intended to capture the potential nonlinearity in the rate of change of differential mortality by education over age. $u_{\text{location:super region}}$ are random intercepts on location, nested within super-regional random intercepts.

Age-period model

Age-period models were fit on all values of $P_{Q,A,S,Y,L}{}''$, which reflect the adjusted input data after cohort extrapolation, to interpolate data for observed cohorts, and to extrapolate to all parts of the desired time series, producing $P_{Q,S,Y,L}{}'''$, single-year estimates of attainment from 1970 to 2018. Several linear mixed-effects models were used and evaluated using out-of-sample predictive validity (see Supplementary Information). Separately for each sex, and region grouping used in the GBD study, the quantity of interest of the country–age–year-specific population, $P_{Q,A,S,Y,L}{}'''$ was estimated:

$$\text{logit}(P_{Q,A,S,Y,L}'') = \beta_{s,r} + \delta_{s,r}\text{year} + I_{s,r} + \alpha_{c,a,s}$$

in which $\beta_{s,r}$ is a sex- and region-specific intercept; $\delta_{s,r}$ captures the linear secular trend for each sex and region; $I_{s,r}$ is a natural spline on age to capture the nonlinear age pattern by sex and region, with knots at 45 and 65 years of age; and $\alpha_{c,a,s}$ is a country-sex-specific random intercept.

Gaussian process regression

Gaussian process regression (GPR) was used to ensure final model results are consistent with input data and incorporate model and data uncertainty to produce uncertainty intervals. GPR has been used extensively as a data synthesis tool[34]. GPR uses a covariance function to smooth the residuals from the age-period model, taking into account the uncertainty in each data point. GPR also synthesizes both data and model uncertainty, in order to produce estimate uncertainty intervals. GPR assumes that the trend in the underlying data follows a Gaussian process, which is defined using a mean function $m(\cdot)$ and a covariance function $\text{Cov}(\cdot)$. Therefore, separately for each Q quantity being estimated, the location–sex–age–year-specific outcome measures are defined:

$$\text{logit}(y_{Q,L,S,C,A}) = g_{Q,L,S,A,Y} + \epsilon_{Q,L,S,A,Y}$$

Where the error term is normally distributed:

$$\epsilon_{Q,L,S,A,Y} = \text{normal}(0, \sigma_p^2)$$

The error variance, σ_p^2 is composed of the squared standard error of the observed data point, as well as the prediction errors from the age-cohort imputation process. The mean function of the model is defined as the age-period model predictions, as detailed above. The covariance function of the model is derived using a Matérn covariance function, consistent with prior applications of GPR:

$$M(y,y') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{d(y,y')\sqrt{2\nu}}{l} \right)^{\nu} K_{\nu} \left(\frac{d(y,y')\sqrt{2\nu}}{l} \right)$$

where $d(\cdot)$ is a distance function, σ^2 is the marginal variance, ν is a smoothness hyper parameter defining the differentiability of the function, l is a link-scale parameter approximately equivalent to the number of years at which two points are no longer correlated, K_{ν} is the Bessel function, and $\Gamma(\cdot)$ is the gamma function. Similar to previous applications of GPR, we approximate σ_p^2 as the super-region and sex-specific residual from the mean function, with ν set to 2 and l to 40, to reflect the inherent smoothness of educational attainment trends over time.

Ensemble *K*-nearest neighbours distribution model

To create a full time-series of distributions of single-years of educational attainment to 2018, we used a K-nearest neighbours algorithm to reconstruct an ensemble distribution for each location–age–sex–year (LASY) combination. To pick K candidate distributions for each LASY combination, we used two modelled entities produced by the above methods, mean educational attainment and proportion of the LASY population with 0 years of schooling, to find the most similar distributions in our database of 3,180 surveys and censuses. The metric used to find the most similar distributions was the Mahalanobis distance:

$$D_M^i(H^i) = \sqrt{(H^i - I^{\text{LASY}})^T S^{-1} (H^i - I^{\text{LASY}})}$$

in which H^i is a multivariate vector $\left(\text{logit}\left(\frac{\text{mean}^i}{18} \right), \text{logit}(\text{prop}_0^i) \right)$ corresponding to a survey–age–sex–year entry in our educational database, I^{LASY} is a multivariate vector $\left(\text{logit}\left(\frac{\text{mean}^{\text{LASY}}}{18} \right), \text{logit}(\text{prop}_0^{\text{LASY}}) \right)$ representing the modelled entities described above, and S^{-1} is the covariance matrix between vectors $\text{logit}\left(\frac{\text{mean}^i}{18} \right)$ and $\text{logit}(\text{prop}_0^i)$.

For each I^{LASY}, K distributions with the smallest Mahalanobis distances are chosen as candidate distributions for the final ensemble distribution. To collapse K distributions to a final ensemble distribution, we use a weighted average of the candidate distributions based on a location, age, and cohort distance defined as:

$$\text{Distance}^i = (P_{\text{age}} \times \text{Distance}_{\text{age}}^i)^{\psi} + (P_{\text{cohort}} \times \text{Distance}_{\text{cohort}}^i)^{\psi} + (P_{\text{space}} \times \text{Distance}_{\text{location}}^i)^{\psi}$$

All values of P and Distance are rescaled to lie between 0.001 and 1. $\text{Distance}_{\text{location}}^i$ is 0.001 for same country, 0.33 for same region, 0.66 for same super-region, and 1 otherwise.

$$\text{Weights}^i = \frac{1}{\text{Distance}^i}$$

ψ is a hyperparameter controlling how sharply weights decrease as Distance^i increases. To collapse K distributions to a final ensemble distribution for each LASY combination we calculated:

$$\text{Proportion}_{\text{eduyrs}}^{\text{LASY}} = \frac{\sum_{i=1}^{K} \text{Weights}^i \times \text{proportion}_{\text{eduyrs}}^i}{\sum_{i=1}^{K} \text{Weights}^i}$$

in which $\text{Proportion}_{\text{eduyrs}}$ is the proportion in each educational bin, 0–18.

Final ensemble distributions were then smoothed by bin using a Loess smoother with a span of η over time to ensure plausible time series for each draw. All hyperparameters were optimized using out-of-sample predictive validity (detailed in the Supplementary Information), and chosen values include: $K = 80$; $P_{\text{age}} = 0.25$; $P_{\text{cohort}} = 0.85$; $P_{\text{space}} = 0.7$; $\psi = 2.5$; $\eta = 0.5$.

Rate of change distribution forecasting model

To forecast the distribution of education and mean years of schooling, we use a rate of change (ROC) model at the single-year bin level. This has the benefit of producing projections of mean attainment that respect the nonlinear dynamics of distributional growth. The model is fit in a timeseries-specific fashion, separately by sex and country. For each single-year bin, we derive a ROC using a weighted average of the ROC for the last 15 years:

$$\text{ROC}_{\text{eduyr}}^{\text{LAS}} = \sum_{i=2004}^{2018} \frac{\text{logit}(\text{proportion}_{\text{eduyr}})^{\text{LAS}^i} - \text{logit}(\text{proportion}_{\text{eduyr}})^{\text{LAS}^{i-1}}}{15}$$

Where $\text{ROC}_{\text{eduyr}}^{\text{LAS}}$ is the average rate of change over the last 15 years within each location–age–sex (LAS) combination for each single-year bin of education (0–18).

The ROC model was leveraged only where the cohort extrapolation model could not inform our estimates. This begins in 2019 for 25–29-year-olds, 2024 for 30–34-year-olds, and 2029 for 35–40-year-olds. For the results presented in the main text, for 25–29-year-olds, this method was used for 2019 onwards.

SDG progress and inequality metrics

Drawing on these estimates of the distribution of years of schooling, we calculate several metrics detailing global progress towards the SDG 4 targets. We calculate the proportion of the population of individuals age 25–29 who have completed primary, secondary, and tertiary education, defined as completing at least 6, 12, and 15 years of schooling, respectively. We describe gender equality using the ratio of female to

male attainment of primary and secondary education, as well as the gap in mean years of schooling between men and women. Aggregate measures at the national level for both sexes, and at the regional level were calculated, using projected population estimates drawn from the World Population Prospects dataset[35]. We also present a novel index of educational inequality among young people in each country, the average AID. This index is defined as the average value of the absolute differences between all possible pairs of individuals in the population. The AID is also mathematically equivalent to the Gini coefficient, multiplied by two times the mean of the distribution[36].

Predictive validity

The main aims of this analysis are predictive in nature, and we therefore assessed each stage of our model, and each model selection decision, with respect to predictive capacity. We focused mainly on 'out-of-sample' predictive ability, which reflects how well the model predicts data that was not directly available. This most mimics the true task that we want our model to accomplish, that is, to make accurate predictions for the geographies and time periods that do not have input data available. To assess out-of-sample predictive validity, we followed the general strategy of dividing our database into 'training' and 'testing' data. The model was fit on the training data, and the results were compared with the testing data. The 'error' of the model represents the average amount that our model was incorrect compared with the 'true' data that was held out. Each step of the modelling process was assessed for how well it predicted (out-of-sample) the mean years of schooling for a given population, as well as other aspects of the distribution, such as the proportion with 0 years of schooling. We also assessed the degree to which predictive validity varied by time period, across regions, and by which type of data source was held out. There were small differences in predictive validity across these dimensions, for example, models tended to perform slightly better in the 2000–2018 period where the most data are available; however, they were generally modest. Furthermore, we found that the best performing models tended to perform optimally across almost all geographies/time periods, so it was not necessary to use multiple models for a single step. All predictive validity results, and a discussion of their implications, can be found in the Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

This study used data that are available from public online repositories, most of which require a straightforward registration process and usage agreement with the data provider. A detailed table of data sources and availability can be found in the Supplementary Information. Although the authors are restricted from providing the data directly in most cases, specific datasets may be made available by request and with permission from the data provider. The authors may be contacted for assistance in acquiring data for the replication of this study. All maps presented in this study have been produced by the authors and no permissions are required for publication. Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset[37].

Code availability

All code used for these analyses is available here: https://github.com/Joseph-Friedman/education_inequality.

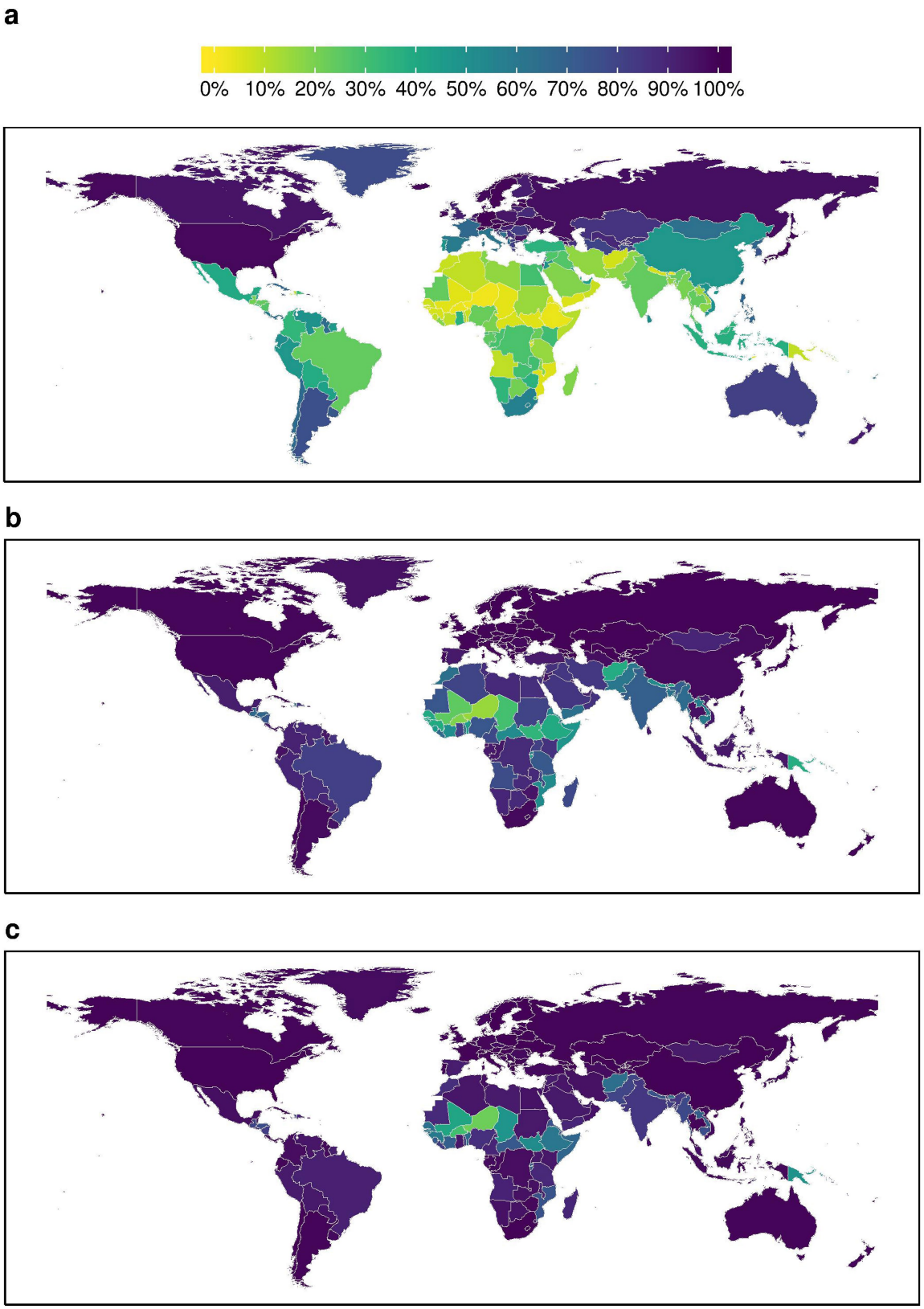
29. Stevens, G. A. et al. Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. *PLoS Med*. **13**, e1002056 (2016).
30. Foreman, K. J. et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. *Lancet* **392**, 2052–2090 (2018).
31. Friedman, J., Graetz, N. & Gakidou, E. Improving the estimation of educational attainment: new methods for assessing average years of schooling from binned data. *PLoS One* **13**, e0208019 (2018).
32. IPUMS International. *YRSCHOOL*. https://international.ipums.org/international-action/variables/YRSCHOOL#comparability_section (accessed November 2018).
33. Barro, R. J. & Lee, J. W. A new data set of educational attainment in the world, 1950–2010. *J. Dev. Econ*. **104**, 184–198 (2013).
34. GBD 2016 Mortality Collaborators. Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1084–1150 (2017).
35. United Nations. *World Population Prospects*. https://population.un.org/wpp/ (accessed November 2018).
36. Gakidou, E. *Health Inequality: Definition, Measurement, and Determinants* (Harvard Univ., 2001).
37. GeoNetwork. *Global Administrative Unit Layers (GAUL)*. http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691 (2007).

Acknowledgements This work was primarily supported by grant OPP1152504 from the Bill & Melinda Gates Foundation. J.F. received support from the UCLA Medical Scientist Training program (NIH NIGMS training grant GM008042). We thank S. B. Munro for assisting with the preparation of the manuscript.

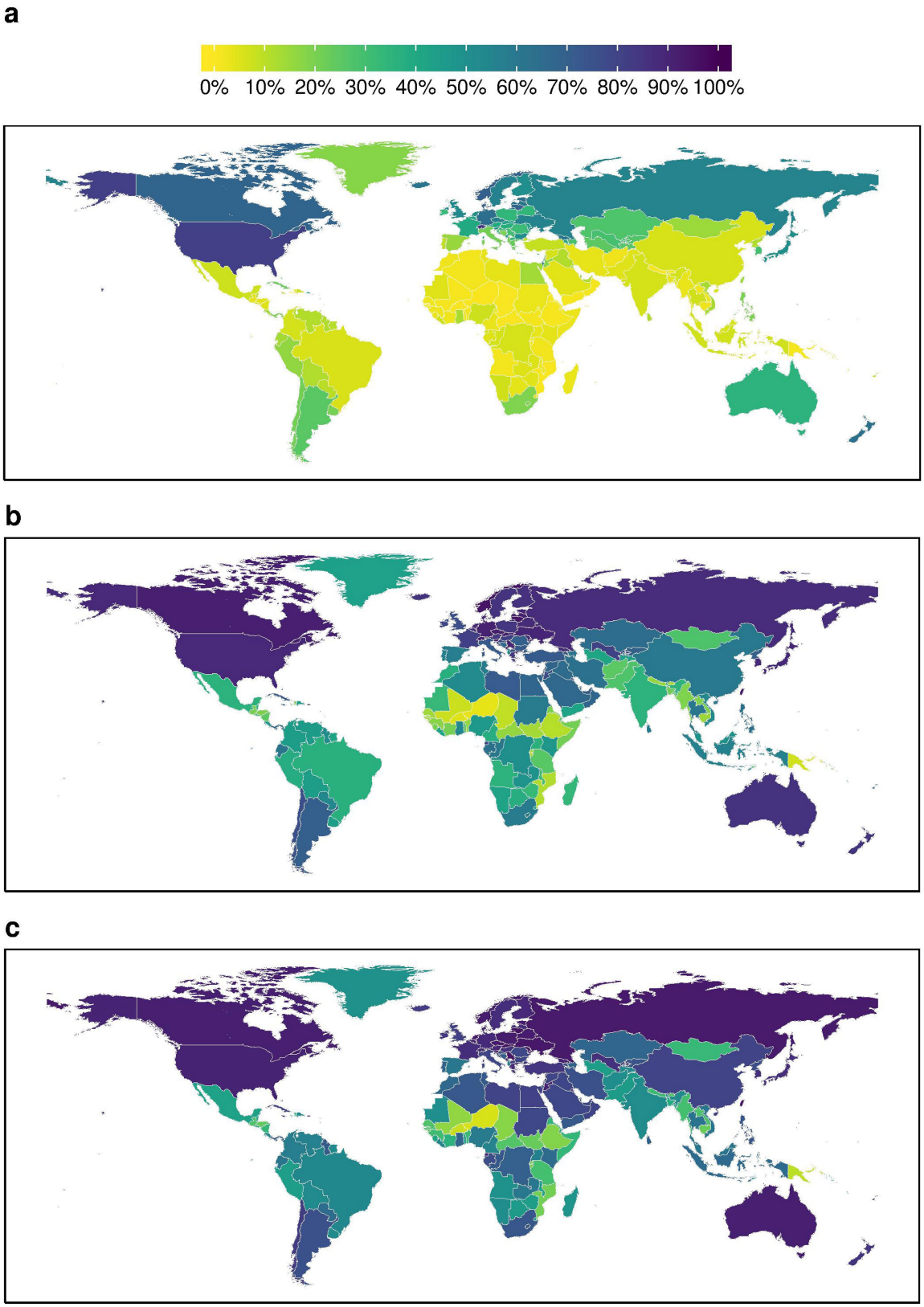
Author contributions J.F., H.Y., N.G. and E.G. conceived and planned the study. J.F., H.Y. and J.W. obtained, extracted and processed educational attainment data. J.F. and H.Y. wrote the computer code and designed and carried out the statistical analyses, with substantial intellectual and methodological inputs from E.G., N.G., L.W. and S.I.H. J.F., H.Y. and E.G. wrote the first draft of the manuscript and all authors contributed to subsequent revisions.

Competing interests The authors declare no competing interests.

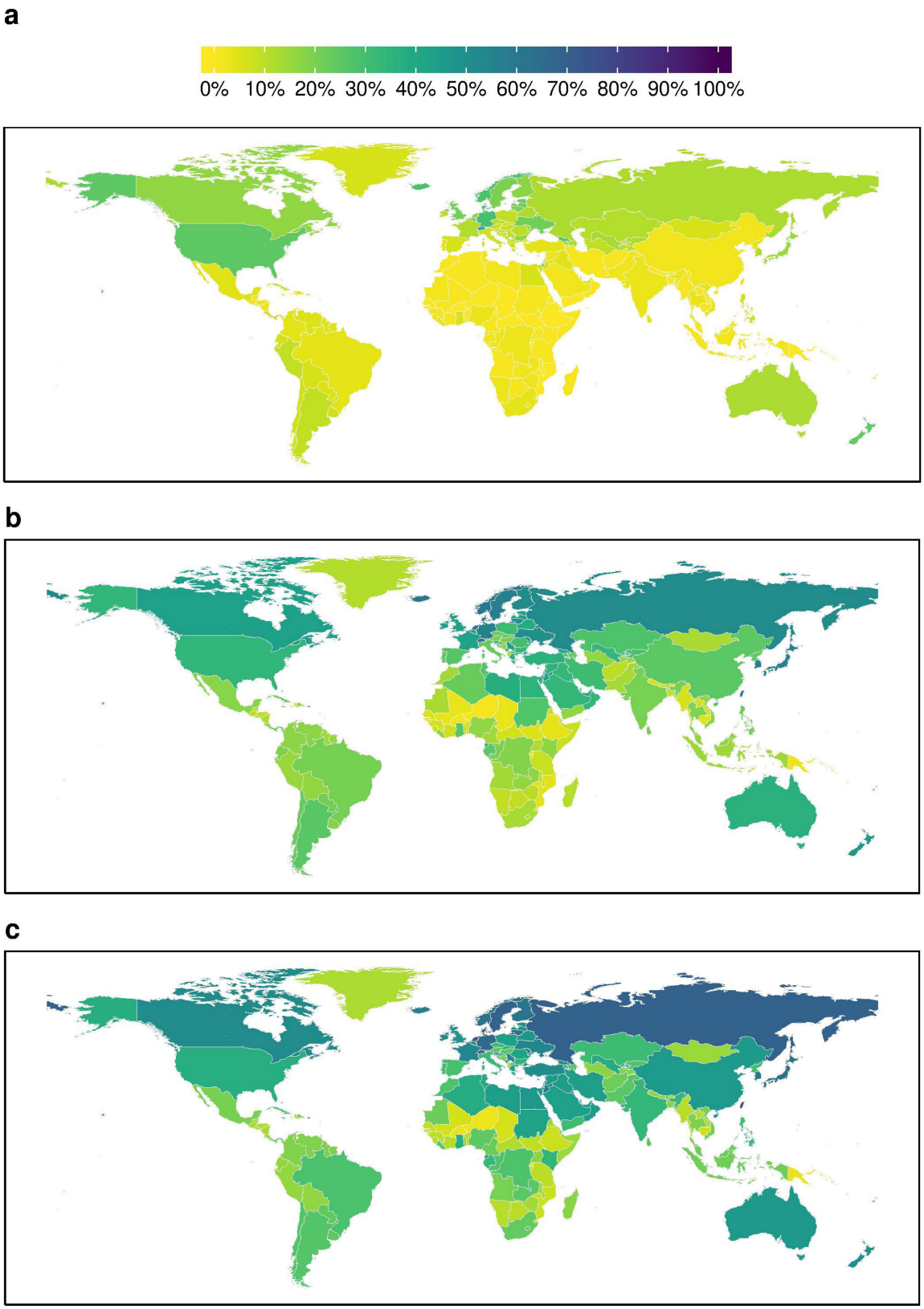
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2198-8.
Correspondence and requests for materials should be addressed to E.G.
Peer review information *Nature* thanks Noam Angrist and Monica Grant for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



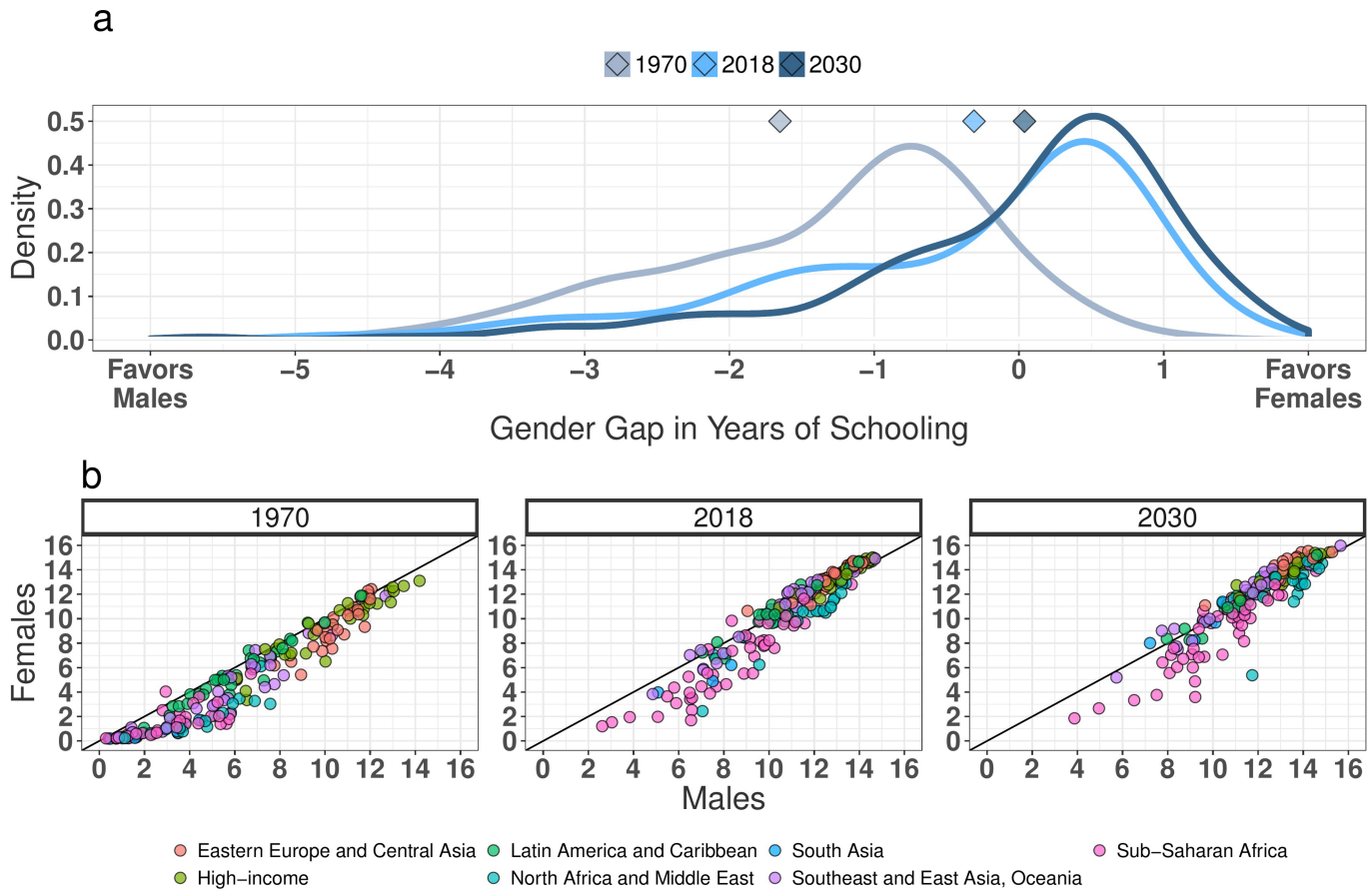
Extended Data Fig. 1 | Completion of 6 or more years of schooling. a–c, The percentage of the population aged 25–29 completing at least 6 years of schooling is shown by country, for 1970 (**a**), 2018 (**b**), and 2030 (**c**). Maps were produced using R v.3.5.0.



Extended Data Fig. 2 | Completion of 12 or more years of schooling. a–**c**, The percentage of the population aged 25–29 completing at least 12 years of schooling is shown by country, for 1970 (**a**), 2018 (**b**), and 2030 (**c**). Maps were produced using R v.3.5.0.



Extended Data Fig. 3 | Completion of 15 or more years of schooling. a–**c**, The percentage of the population aged 25–29 completing at least 15 years of schooling is shown by country, for 1970 (**a**), 2018 (**b**), and 2030 (**c**). Maps were produced using R v.3.5.0.



Extended Data Fig. 4 | Years of schooling among men and women. a, The distribution of the gap in mean years of schooling between men and women, aged 25–29, is shown for 1970, 2018, and 2030, with the population-weighted mean for each time point represented with a diamond. Means were calculated over modelled estimates from $n = 195$ countries. **b**, Years of schooling is represented for men on the *x* axis and women on the *y* axis for 1970, 2018, and 2030, in which each point indicates the value for one country, colour-coded by regional grouping. A point above the line indicates additional schooling for women relative to their male counterparts.

nature research

Corresponding author(s): Emmanuela Gakidou

Last updated by author(s): Feb 4, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No primary data collection was carried out for this analysis.
Data analysis	All analyses were conducted using R version 3.1.3 and Python 2.7.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

This study used data that are available from public online repositories, but which in most cases require a straightforward registration process and usage agreement with the data provider. A detailed table of data sources and availability can be found in the supplement. Although the authors are restricted from providing the data directly in most cases, specific data sets may be made available by request and with permission from the data provider. The authors may be contacted for assistance in acquiring data for the replication of this study.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A descriptive, population-level, ecological study of inequality
Research sample	All available nationally representative census and survey data containing information about educational attainment.
Sampling strategy	We used all available nationally representative census and survey data containing information about educational attainment. .
Data collection	N/A - secondary analysis
Timing	N/A - Secondary analysis, data originally collected between 1950 and 2018 were included.
Data exclusions	As described in the methods section, with greater detail in the supplement, this study provides modeled estimates of the single-year distribution of years of schooling over time and by country. Data were included from 195 nations and territories that are part of the Global Burden of Disease 2017 study. Data for other areas that do not pertain to this list, or which were found to not be nationally representative, were not included. Data that did not include 5-year age groups, or which were not disaggregated by age or sex were also not included. Data from outside the 1950-2018 time period were also not included.
Non-participation	N/A - secondary analysis of nationally-representative statistics
Randomization	N/A - observational analysis, no experimental groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

The mutational landscape of normal human endometrial epithelium

https://doi.org/10.1038/s41586-020-2214-z

Received: 19 December 2018

Accepted: 20 March 2020

Published online: 22 April 2020

Check for updates

Luiza Moore[1,2], Daniel Leongamornlert[1], Tim H. H. Coorens[1], Mathijs A. Sanders[1,3], Peter Ellis[1,4], Stefan C. Dentro[1,5], Kevin J. Dawson[1], Tim Butler[1], Raheleh Rahbari[1], Thomas J. Mitchell[1], Francesco Maura[1,6], Jyoti Nangalia[1], Patrick S. Tarpey[1], Simon F. Brunner[1], Henry Lee-Six[1], Yvette Hooks[1], Sarah Moody[1], Krishnaa T. Mahbubani[7,8,9], Mercedes Jimenez-Linan[2], Jan J. Brosens[10], Christine A. Iacobuzio-Donahue[11,12], Inigo Martincorena[1], Kourosh Saeb-Parsy[7,8], Peter J. Campbell[1] & Michael R. Stratton[1]✉

All normal somatic cells are thought to acquire mutations, but understanding of the rates, patterns, causes and consequences of somatic mutations in normal cells is limited. The uterine endometrium adopts multiple physiological states over a lifetime and is lined by a gland-forming epithelium[1,2]. Here, using whole-genome sequencing, we show that normal human endometrial glands are clonal cell populations with total mutation burdens that increase at about 29 base substitutions per year and that are many-fold lower than those of endometrial cancers. Normal endometrial glands frequently carry 'driver' mutations in cancer genes, the burden of which increases with age and decreases with parity. Cell clones with drivers often originate during the first decades of life and subsequently progressively colonize the epithelial lining of the endometrium. Our results show that mutational landscapes differ markedly between normal tissues—perhaps shaped by differences in their structure and physiology—and indicate that the procession of neoplastic change that leads to endometrial cancer is initiated early in life.

Acquisition of mutations is a ubiquitous feature of cells in living organisms. Although there has been comprehensive characterization of the somatic mutation landscape of human cancer[3–5], knowledge of the patterns of somatic mutation in normal cells is limited. This has mainly been due to the challenge of detecting somatic mutations in normal tissues. Several strategies have recently been developed to address this, including the sequencing of in vitro-derived cell clones from normal tissues[6–8], the sequencing of small biopsies that contain limited numbers of microscopic clones[9–12], the sequencing of microscopically distinguishable structural elements that are clonal units[13–15], highly error-corrected sequencing[16,17] and the sequencing of single cells[18]. Together, these approaches have begun to reveal differing mutation burdens between cell types, the patterns of acquisition of mutation burdens over time and the underlying mutational processes. These strategies have also shown that clones of normal cells with driver mutations in cancer genes are present in normal tissues. In the glandular epithelium of the colon, these mutations are relatively uncommon[14]—but in the squamous epithelia of the skin[9] and oesophagus[10], and in the blood[19–21], clones that carry drivers can constitute substantial proportions of the normal cells present after middle age.

The factors that determine differences in the mutation landscape between normal cell types are incompletely understood. However, these factors plausibly include the intrinsic structural and physiological features of each tissue. The endometrium is a uniquely dynamic tissue composed of a stromal cell layer invaginated by a contiguous glandular epithelial sheet that covers the luminal surface. Endometrium adopts multiple different physiological states during life, including in premenarche, menstrual cycling, pregnancy and postmenopause. During reproductive years, the endometrium undergoes cyclical breakdown, shedding, repair and remodelling in response to oscillating levels of oestrogen and progesterone, which together entail the iterative restoration of the contiguity of the interrupted glandular epithelial sheet that is effected by stem cells within basal glands retained after menstruation[1,2,22].

The characterization of the mutational landscapes of normal tissues is advancing our understanding of the succession of intermediate neoplastic stages between normal cells and the cancers that originate from them. Endometrial cancer is the most common gynaecological tumour in high-income countries, with a peak incidence at 75–80 years of age[23]. There are two major histological classes[24,25]. Type I, endometrioid carcinoma, is the more common of the two; the main known risk factor is oestrogen exposure, influenced by ages of menarche and menopause, and body mass index[24,26]. Type II, which includes serous and clear cell carcinomas, occurs in older women, with smoking and

[1]Cancer, Ageing and Somatic Mutation (CASM), Wellcome Sanger Institute, Cambridge, UK. [2]Department of Pathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [3]Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands. [4]Inivata Ltd, Cambridge, UK. [5]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [6]Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [7]Department of Surgery, University of Cambridge, Cambridge, UK. [8]Cambridge NIHR Biomedical Research Centre, Cambridge, UK. [9]Department of Haematology, University of Cambridge, Cambridge, UK. [10]Tommy's National Miscarriage Research Centre, Warwick Medical School, University of Warwick, Coventry, UK. [11]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [12]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ✉e-mail: mrs@sanger.ac.uk

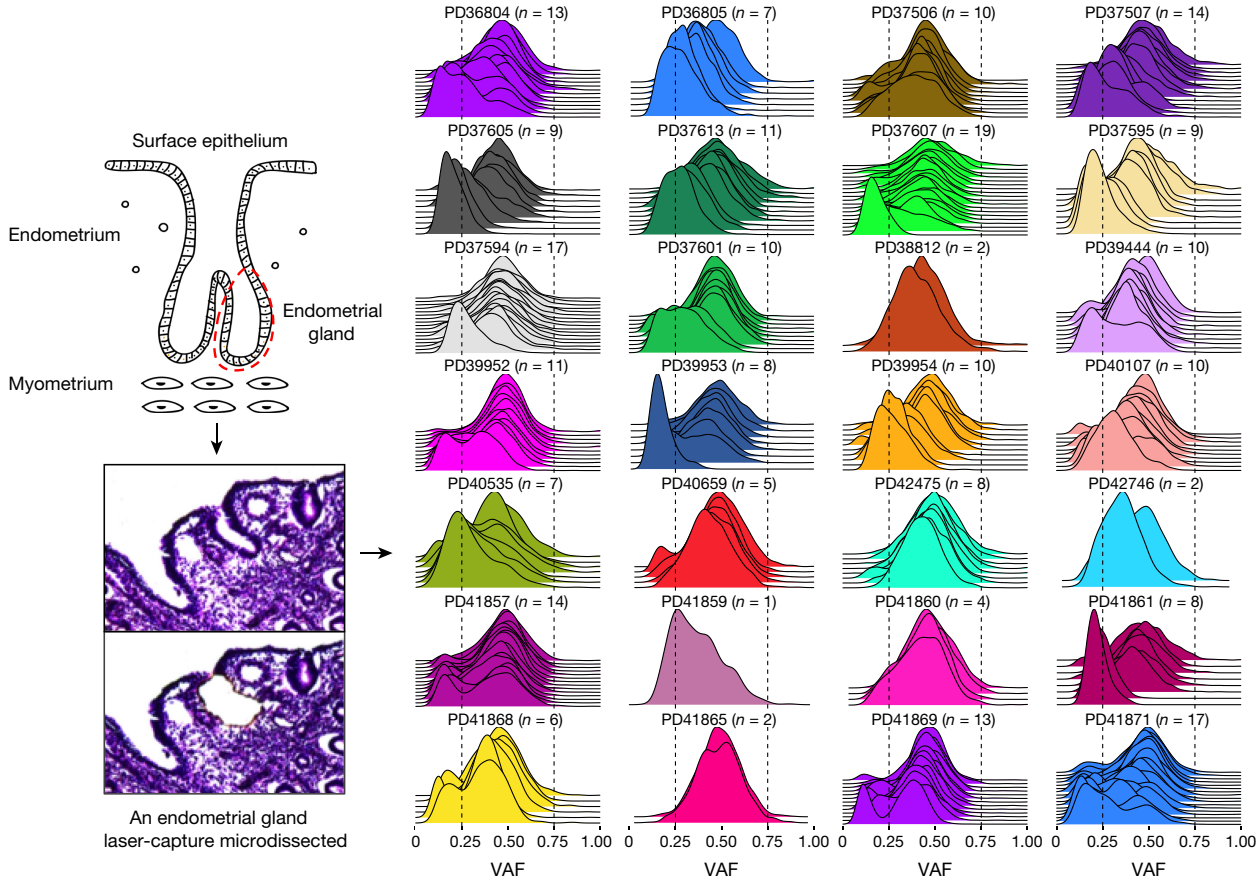


Fig. 1 | Clonality of normal endometrial glands. Individual normal endometrial glands were laser-capture microdissected and whole-genome sequenced. Most (91%, 234 out of 257) of the glands were clonal cell populations with a median VAF between 0.3 and 0.5 for base substitutions. Each density line represents an endometrial gland sample; individual samples are grouped and coloured by patient ($n = 28$).

body mass index as risk factors[27]. Commonly mutated cancer genes include *PTEN*, *TP53*, *PIK3CA*, *KRAS*, *ARID1A*, *FBXW7* and *PIK3R1*[28], and subsets of endometrial cancer carry many base substitution and/or small insertion and deletion (indel) mutations due to defective DNA mismatch repair or polymerase proof-reading mutations, or many copy number changes and genome rearrangement[29].

Recent studies using targeted sequencing have revealed driver mutations in known cancer genes in a high proportion of endometrial glands in endometriosis[13,30,31] and eutopic normal endometrial epithelium[13,32]. Here, by whole-genome sequencing of individual glands, we comprehensively characterize the mutational landscape of normal endometrial epithelium, explore the influences of age and parity, and estimate the timing of driver mutations.

Samples and sequencing

We used laser-capture microdissection to isolate 292 histologically normal endometrial glands from 28 women aged between 19 and 81 years. Samples were obtained from biopsies taken for the investigation of reproductive problems (14 women), hysterectomies for benign non-endometrial pathologies (2 women), residual tissues from transplant organ donors (8 women) and autopsies after death from nongynaecological causes (4 women). DNA from each gland was whole-genome sequenced using a protocol that accommodates small amounts of input DNA[14]. The mean sequencing coverage was 28-fold; only samples with >15-fold coverage were included in subsequent analyses ($n = 257$) (Supplementary Results 1, 2). Somatic mutations in each gland were determined by comparison with whole-genome sequences from other tissues from the same individuals.

Clonality of endometrial glands

To assess whether endometrial glands comprise clonal cell populations, we examined the variant allele fractions (VAFs) of somatic mutations. Ninety-one per cent (234 out of 257) of microdissected endometrial glands showed distributions of VAFs with peaks between 0.3 and 0.5 (Fig. 1, Extended Data Fig. 1a), indicating that each gland consists predominantly of a cell population that is descended from a single epithelial progenitor stem cell (a formal clonality analysis is described in Methods, Extended Data Fig. 2, Supplementary Results 3). Subsequent analyses (described in 'Driver mutations') revealed that many endometrial glands carry driver mutations in known cancer genes. However, endometrial glands exhibited clonality irrespective of the presence of driver mutations (Extended Data Fig. 1b, Supplementary Results 4). Thus, colonization of endometrial glands by descendants of single endometrial epithelial stem cells is not contingent on a selective growth advantage provided by driver mutations, and may occur by a process analogous to genetic drift (as previously proposed for other tissues[33,34]).

Mutation burdens and signatures

Somatic mutation burdens in normal endometrial glands from the 28 women ranged from 209 to 2,833 base substitutions (median of 1,521) and 1 to 358 indels (median of 180) (Fig. 2a, b). This variation was predominantly attributable to age, with about 29 base substitutions per gland per year being acquired during adult life (linear mixed-effect model, 95% confidence interval 23–34, $P = 3.02 \times 10^{-11}$) (Supplementary Results 5, 6). The presence of a driver mutation was also associated with

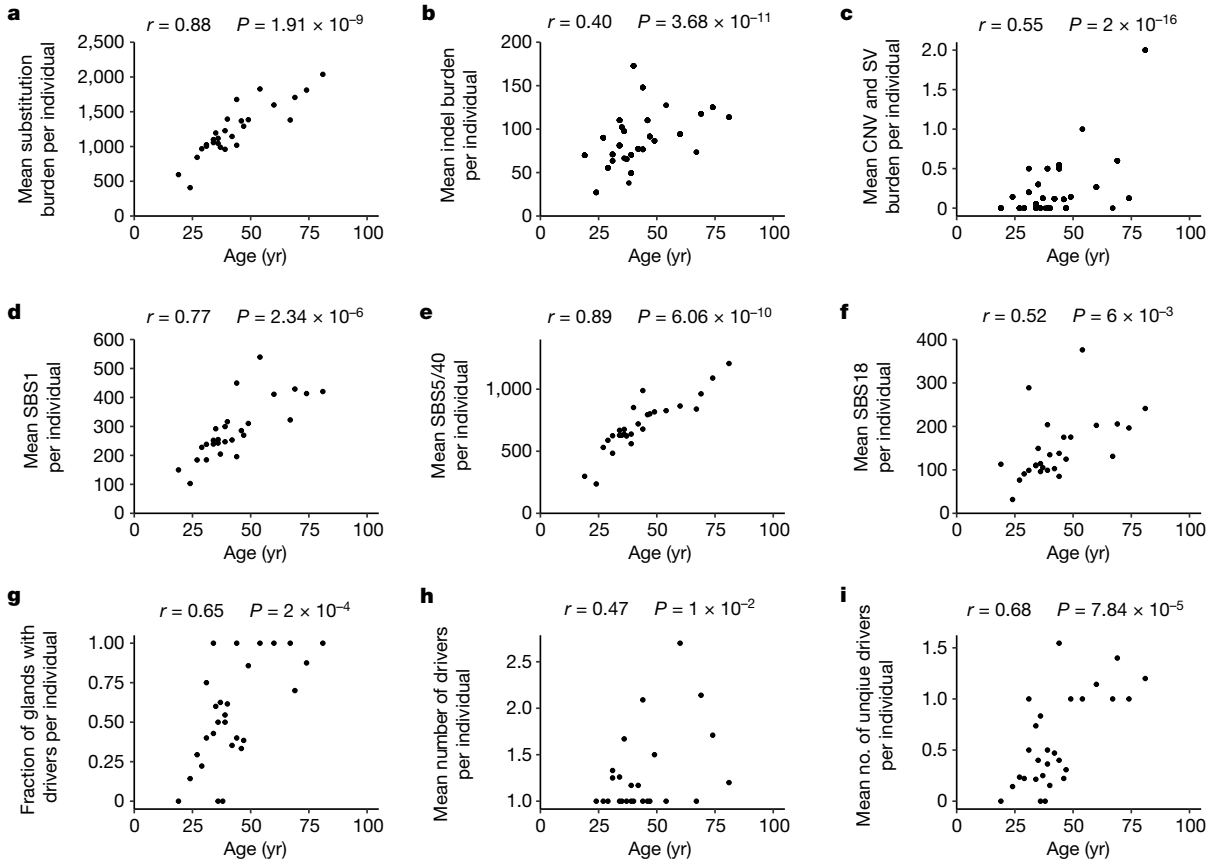


Fig. 2 | Mutation burden correlates with age in normal endometrial glands. Mutation burdens shown as mean for each donor ($n = 28$ donors), with Pearson correlation (r) with age and P values (P) from linear regression (burden–age). **a**–**c**, Variant burdens. **a**, Substitution burden. **b**, Indel burden. **c**, Copy-number variant (CNV) and structural variant (SV) burden. **d**–**f**, SBS burdens. **d**, SBS1 burden. **e**, SBS5/40 burden. **f**, SBS18. **g**–**i**, Driver mutation burden per gland. **g**, Fraction of glands with drivers, per individual. **h**, Mean number of driver mutations in glands with drivers. **i**, Mean number of unique (different) driver mutations per gland.

an additional approximately 110 substitutions (95% confidence interval 43–177, $P = 1.34 \times 10^{-3}$). There was no obvious correlation between parity and total somatic mutation burden.

We identified five previously described single-base-substitution (SBS) mutational signatures (Supplementary Results 7–9): SBS1, which is predominantly characterized by NCG > NTG mutations and is probably due to spontaneous deamination of 5-methylcytosine; SBS5 and SBS40, two relatively featureless 'flat' signatures of uncertain cause; SBS18, predominantly characterized by C > A substitutions and possibly due to reactive oxygen species[35]; and SBS23, a signature predominantly composed of C > T mutations and of unknown aetiology. Because SBS5 and SBS40 are relatively featureless, it is challenging to estimate their separate contributions[4] and they have therefore been combined (designated SBS5/40) (but shown separately in Supplementary Results 8, 9). SBS23 has previously occasionally been found in liver cancers with high mutation burdens. Given the low mutation burden and small contribution of SBS23 in the data reported here, it is unclear whether this is the same signature and so SBS23 was included in the 'unattributable' category. The mean signature exposures were 0.23 for SBS1, 0.58 for SBS5/40 and 0.12 for SBS18. There were positive linear correlations with age for the mutation burdens attributable to each of these three signatures (Fig. 2d–f). To ascertain the periods during which different mutational processes operate, we constructed phylogenetic trees of endometrial glands for each individual, which indicated that the mutational processes that underlie these three signatures are active throughout life (Figs. 3, 4, Extended Data Fig. 3). In regard to small indels, single T insertions at runs of T bases were the most common type of mutation that we observed (Supplementary Results 10).

Somatic copy-number changes and structural variants were found in 36 out of 257 (14%) normal endometrial glands, almost all of which carried just a single change (Extended Data Fig. 4, Supplementary Results 4). These changes included copy-number neutral loss of heterozygosity in 8 glands, whole chromosome copy-number increase in 1 gland and structural variants in 18 glands (12 large deletions, 6 tandem duplications and 9 translocations). One of three glands carrying a *TP53* mutation exhibited nine structural variants, indicating that genomic instability caused by defective DNA maintenance occurs in normal cells.

Driver mutations

To identify genes under positive selection, we used a statistical method based on the observed:expected ratios of nonsynonymous:synonymous mutations[28]. Twelve genes showed evidence of positive selection in the 257 normal endometrial glands: *PIK3CA*, *PIK3R1*, *ARHGAP35*, *FBXW7*, *ZFHX3*, *FOXA2*, *ERBB2*, *CHD4*, *KRAS*, *SPOP*, *PPP2R1A* and *ERBB3* (Supplementary Results 11). All were listed among 369 genes that have previously been shown to be under positive selection in human cancer[28]. To identify additional drivers in the 257 endometrial glands, we sought mutations with the characteristics of drivers in those 369 genes (Methods). In total, we found 209 driver mutations in normal endometrial glands from 25 out of 28 women (Supplementary Results 4). The youngest carrier was a 24-year-old woman (patient PD40535) with a $KRAS^{G12D}$ mutation in 1 out of 7 glands that we sampled. We found that 147 out of 257 endometrial glands carried at least 1 driver mutation; 42 out of 257 glands carried at least 2 drivers; and 5 out 257 glands carried at least 4 drivers. In 4 women (aged 34 (19 glands), 44 (11 glands),

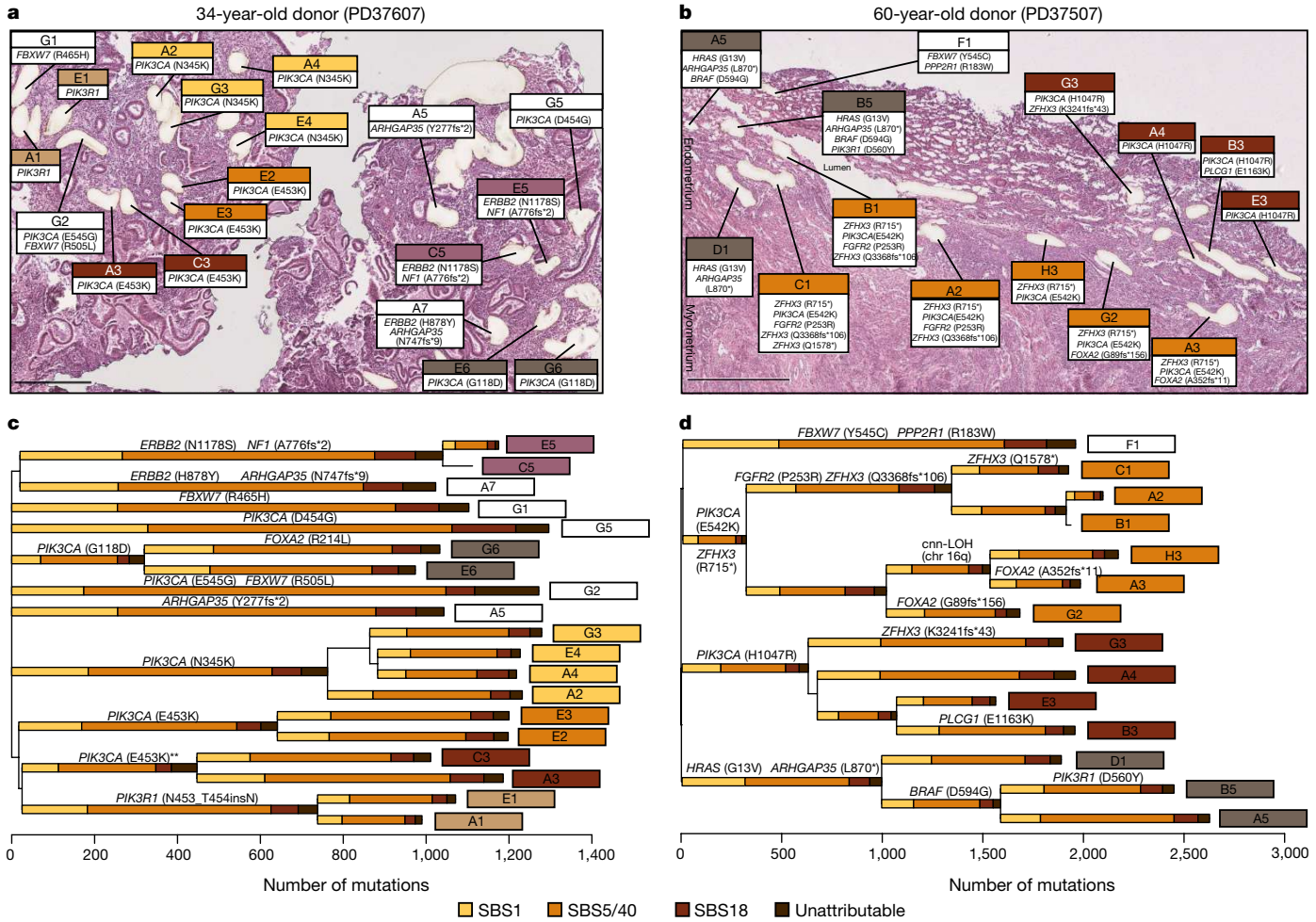


Fig. 3 | Histology images and reconstructed phylogenetic trees for two individuals in whom every normal endometrial gland contained at least one driver mutation. a, **b**, Haematoxylin and eosin images of endometrial glands from a 34-year-old woman (**a**) and a 60-year-old woman (**b**) were taken after laser-capture microdissection (20× magnification). **c**, **d**, Phylogenetic trees were reconstructed for the 34-year-old woman (**c**) and 60-year-old woman (**d**) using SBSs; the length of each branch is proportional to the number of variants. A stacked bar plot of the attributed SBS mutational signatures that contributed to each branch is then superimposed onto every branch; signature extraction was not performed on branches with fewer than 100 substitutions. The ordering of signatures within each branch is for visualization purposes only, as it is not possible to time the different signatures within individual branches. Glands that shared over 100 variants were considered part of the same clade (indicated by the colour of the sample identifier label). Glands that did not belong to any clades are in white. SBS signatures are colour-coded; substitutions that were not attributed to the reference signatures, and those attributed to SBS23, are shown as 'unattributable'. Scale bars, 500 µm.

60 (14 glands) and 81 (5 glands)), all of the glands that we analysed carried driver mutations, which suggests that the whole endometrium had been colonized by microneoplastic clones (Fig. 3, Extended Data Fig. 3). The fraction of endometrial glands carrying a driver (Fig. 2g), the mean number of drivers per gland (Fig. 2h) and the number of different drivers in each individual (corrected for the number of glands sampled) (Fig. 2i) all positively correlated with age of the individual. However, there were sufficient outliers to suggest that other factors influence the colonization of the endometrium by driver-carrying clones. Indeed, our generalized linear mixed-effect model showed that in addition to the positive association of age with the accumulation of driver mutations (0.035 driver mutations per year, 95% confidence interval 0.01–0.06, $P = 3.31 \times 10^{-4}$), parity had a negative association (−0.253 driver mutations per life birth, 95% confidence interval −0.46 to −0.05, $P = 1.33 \times 10^{-2}$) (Supplementary Results 12, 13).

We found driver mutations in recessive (tumour-suppressor genes) and dominant cancer genes, similar to recent publications[13,30,32]. *PIK3CA* was the most frequently mutated cancer gene (Fig. 3, Extended Data Figs. 3, 5, Supplementary Results 14). Most truncating drivers in recessive cancer genes were heterozygous, indicating that haploinsufficiency confers a growth advantage in normal cells. Nevertheless, further inactivating mutations in the same genes in other glands show that an additional advantage is conferred by complete abolition of their activity (notably for *ZFHX3* in the 60-year-old woman) (Fig. 3). Driver mutations were found in genes that encode growth factor receptors (*ERBB2*, *ERBB3* and *FGFR2*), components of signal transduction pathways (*HRAS*, *KRAS*, *BRAF*, *PIK3CA*, *PIK3R1*, *ARHGAP35*, *RRAS2*, *NF1*, *PPP2R1A* and *PTEN*), pathways that mediate responses to steroid hormones (*ZFHX3*, *FOXA2* and *ARHGAP35*), proteins involved in chromatin function (*KMT2D* and *ARID5B*) and protein-mediated degradation pathways (*FBXW7*) that target oncoproteins, such as mTOR and MYC. Many different combinations of mutated cancer genes were found in individual glands.

Timing of driver mutations

Constructing phylogenetic trees of individual endometrial glands enabled the characterization of the mode of expansion of normal cell clones with drivers and the timing of their initiation. Glands with a phylogenetically close relationship were often in close physical

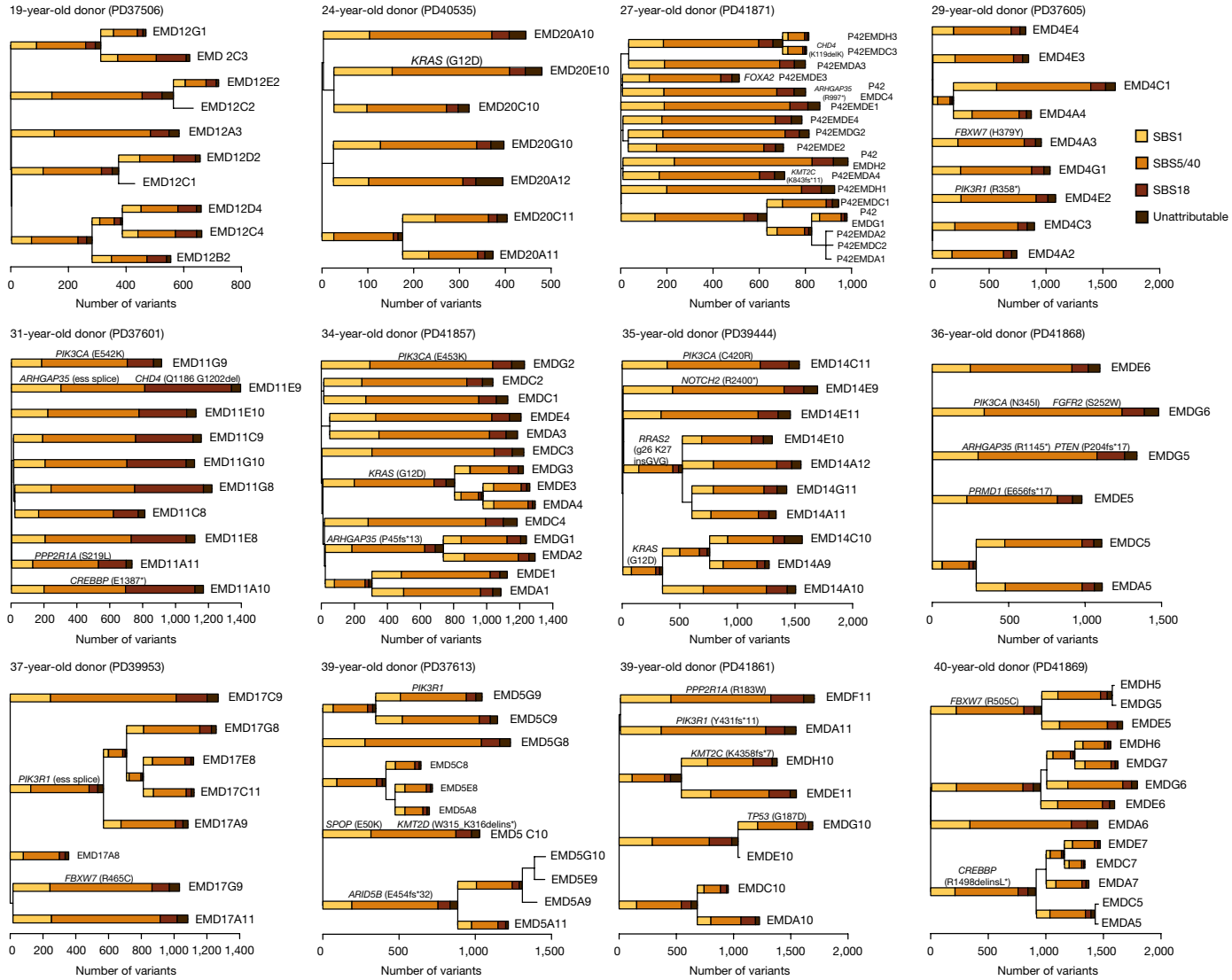


Fig. 4 | Phylogenetic trees of endometrial glands for donors aged 19 to 40 years. Phylogenetic trees for individuals aged 19 to 40 years were reconstructed using SBSs with branch length proportional to the number of variants; the stacked bar plots represent the attributed SBS mutational signatures that contributed to each branch. Signature extraction was not performed on branches with fewer than 100 substitutions. The ordering of signatures within each branch is for visualization purposes only, as it is not possible to time different signatures within individual branches. SBS signatures are colour-coded; substitutions that were not attributed to the reference signatures, and those attributed to SBS23, are shown as 'unattributable'. EMD codes refer to individual endometrial glands.

proximity within the endometrium (Fig. 3). In phylogenetic clusters for which the mutation catalogues were almost identical, this may simply reflect multiple sampling of a single tortuous gland that weaves in and out of the plane of section, rather than distinct glands with their own stem cell populations (for example, glands C5 and E5 in Fig. 3a, c). For other phylogenetic clusters, the different branches within the clade have diverged substantially, sometimes acquiring different driver mutations, and therefore are probably derived from different stem cell populations. In such instances, phylogenetically related glands can range over distances of hundreds of micrometres, which suggests that their clonal evolution has entailed the capture and colonization of extensive zones of the endometrial lining (for example, glands C1, A2, B1, H2, A3 and B3 in Fig. 3b, d). Conversely, some glands in close physical proximity are phylogenetically distant (for example, glands E1 and G2 in Fig. 3a, c), indicating that their cell populations have remained isolated from each other.

Driver mutations were positioned on the phylogenetic trees for each individual, and times of occurrence were estimated by assuming constant somatic mutation rates during life (Fig. 5, Extended Data Figs. 6, 7, Methods). Although this assumption is unlikely to be completely correct, the results show that mutations in normal endometrial cells are acquired in a more-or-less linear fashion throughout life and potential modifying factors, including acquisition of a driver, make only modest differences to mutation rates. Furthermore, overall our approach is likely to overestimate the ages before which driver mutations have occurred, because it does not account for the time taken for a single endometrial stem cell to colonize an individual gland, which—in colorectal crypts—has been estimated to take several years[36]. Therefore, our results indicate that at least some driver mutations occur early in life. These included a *KRAS*^{G12D} mutation in 3 glands from a 35-year-old woman, and a *PIK3CA* mutation in 2 glands from a 34-year-old woman, both of which are likely to have arisen during the first decade of life (Figs. 3, 4, Extended Data Figs. 6, 7). A pair of drivers in *ZFHX3* and *PIK3CA*, which co-occur in 6 glands from a 60-year-old woman, was also acquired during the first decade of life, indicating that driver-associated clonal evolution also begins early in life (Figs. 3, 5). It is possible that

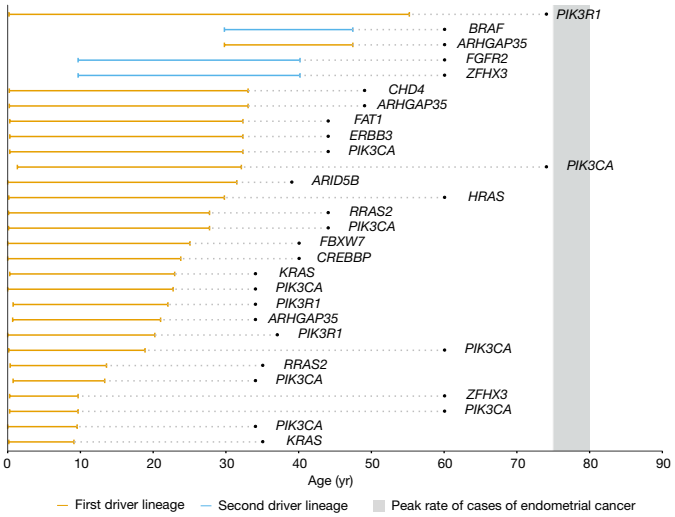


Fig. 5 | Timing of driver mutations in normal endometrial glands. To time the driver mutations, we reconstructed phylogenetic trees using SBSs. To estimate the time interval in which specific mutations occurred, we used two approaches (Methods). We calculated a patient-specific mutation rate by taking the ratio of the mean mutation burden per endometrial gland of the patient and age of the patient. The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. A similar approach was used for timing indels. We timed the driver mutations that occurred in the 'trunks' and branches. We display only those driver variants that occurred in the 'trunks' of the individual trees. We show that many such events occur decades before the reported peak incidence of endometrial cancer (variants with an interval of <1 year between the upper age and the age at sampling were excluded from this plot for illustration purposes). On the basis of our calculations, four driver variants (*KRAS*^{G12D}, *PIK3CA*^{G118D}, *PIK3CA*^{E542K} and *ZFHX3*^{R715*}) from three different women occurred before the age of ten.

many more clones with drivers were initiated during the first decade of life, but their phylogenetic trees are not informative in this regard (Extended Data Figs. 6, 7). However, there was also evidence for the continued accumulation and clonal expansion of driver mutations into the later decades of life (Fig. 5, Extended Data Figs. 6, 7).

Comparison between normal tissue and cancer

Endometrial cancers (from the recent Pan Cancer Analysis of Whole Genomes (PCAWG) dataset[4]) exhibited higher mutation loads than normal endometrial cells for base substitutions (about 5-fold higher, medians of 1,346 and 7,330 in normal endometrium and endometrial cancer, respectively (Mann–Whitney *U*-test, $P = 7.63 \times 10^{-6}$)) and indels (Extended Data Fig. 8a, b). These differences also pertained to normal endometrial cells with driver mutations. In most endometrial cancers, the differences are attributable to higher mutation burdens of the ubiquitous base substitution and indel mutational signatures[3,4]. In addition, however, the very high mutation loads of the subsets of endometrial cancer with deficiencies in DNA mismatch repair and proof-reading mutations in polymerase-ε or polymerase-δ were not seen in normal endometrial cells. Differences between endometrial cancers and normal cells were even more marked for structural variants and copy-number changes (median number zero in normal endometrial cells and about 23 in endometrial cancers[37]), and this difference again pertained to normal endometrial cells with drivers.

There were also differences in the repertoire of cancer genes in which driver mutations were found (Extended Data Fig. 8c–e, Supplementary Results 4, 11). Notably, mutations in *PTEN*, *CTCF*, *CTNNB1* and *ARID1A* in endometrioid, and in *TP53* in serous carcinoma of the endometrium accounted for higher proportions of driver mutations than in normal endometrial cells. It is possible that *PTEN*, *ARID1A*, *TP53* and *CTCF* require biallelic mutation to confer a growth advantage and this may account for their lower prevalence in normal cells. However, heterozygous mutations in *PTEN* and *TP53* were found, albeit only in around 2% (5 out of 257) of all sampled glands, and this explanation would not account for the relative deficit of *CTNNB1* mutations. Overall, the results suggest that driver mutations in some cancer genes are relatively effective at enabling the colonization of normal tissues, but confer a limited risk of conversion to invasive cancers. Conversely, other drivers may require biallelic mutation and/or confer limited advantage in colonizing normal tissues, but are relatively effective at the conversion to malignancy.

Discussion

Studies of normal endometrial epithelium and other types of normal cell[6,7,9,10,13–15,19,20] are revealing the landscape of somatic mutations in normal human cells. Somatic mutations are predominantly generated by a limited repertoire of ubiquitous mutational processes that generate base substitutions, small indels, genome rearrangements and whole chromosome copy-number changes, which exhibit more-or-less constant mutation rates during life. Additional mutational processes present only in some cells, some cell types and/or that are intermittent also contribute to the mutation burden—albeit apparently not in the endometrial epithelium.

The prevalence of clones with driver mutations is substantially different in different types of normal cell. Numerous cell clones with one or more driver mutations colonize much of the normal endometrial epithelium (as discussed in this Article, and in previous studies[13,32]), in contrast to another glandular epithelium, the colon, in which about 1% of normal crypts in middle-aged individuals carry a driver[13,14]. This is unlikely to be due to differences in the somatic mutation rate between endometrial and colonic epithelial cells, which are relatively modest; in any case, the somatic mutation rate is higher in the colon[6,14,38]. However, it may be attributable to intrinsic differences in structure and physiology between the endometrium and colon. In the endometrium, the cyclical process of tissue breakdown, shedding and remodelling iteratively opens up denuded terrain for pioneering clones of endometrial epithelial cells with drivers to preferentially colonize, compared to wild-type cells. In the colon, however, the selective advantage of a clone with a driver is usually confined to the small, siloed population of a single crypt, with only occasional opportunities for further expansion. Although the colonization of endometrium by driver clones progresses with age, it is already well-advanced in some young women—and parity has an inhibitory effect on it. The effect of parity is of particular interest as increased parity reduces the risk of endometrial cancer and it is conceivable that this is mediated by its effect on the expansion of driver clones[39]. Further studies of normal endometrium are required to assess how premenarchical and postmenopausal states, hormone contraceptive use and hormone replacement therapies influence the mutational landscape and its potential effect on pregnancy and fertility.

The burdens of all mutation classes are lower in normal endometrial cells (including those with drivers) than in endometrial cancers. Therefore, in endometrial epithelial stem cells, and in all other tissues studied thus far (including colon, oesophagus and skin), normal mutation rates are sufficient to generate large numbers of clones with driver mutations that behave as normal cells, but acquisition of an elevated mutation rate and burden is associated with further evolution to invasive cancer[9,10,14]. Because the endometrial epithelium is extensively colonized by clones of normal cells with driver mutations in middle-aged women, and the lifetime risk of endometrial cancer is only 3% (ref. [23]), this conversion from a normal cell clone with drivers to symptomatic malignancy appears to be extremely rare.

The first driver mutations in normal endometrial clones with drivers can arise within the first decade of life, and our results are compatible

with of many doing so. The modal period of diagnosis of endometrial cancer is 75–80 years of age. Therefore, if normal cell clones with drivers are progenitors of endometrial cancers (which is plausible given the similar driver mutations found), our results suggest that many cancers are initiated during childhood and evolution to malignancy takes place over the lifetime of an individual. This perspective on the long duration of neoplastic evolution of invasive endometrial cancer has resonance with previous observations on leukaemia[40,41] and, more recently, other solid malignancies[42–45], and may be a common feature of the development of human cancers.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2214-z.

1. Gargett, C. E., Schwab, K. E. & Deane, J. A. Endometrial stem/progenitor cells: the first 10 years. *Hum. Reprod. Update* **22**, 137–163 (2016).
2. Kaitu'u-Lino, T. J., Ye, L. & Gargett, C. E. Reepithelialization of the uterine surface arises from endometrial glands: evidence from a functional mouse model of breakdown and repair. *Endocrinology* **151**, 3386–3395 (2010).
3. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
4. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
5. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
6. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
7. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
8. Franco, I. et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. *Nat. Commun.* **9**, 800 (2018).
9. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
10. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
11. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
12. Zhu, M. et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. *Cell* **177**, 608–621 (2019).
13. Suda, K. et al. Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. *Cell Rep.* **24**, 1777–1789 (2018).
14. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
15. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
16. Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).
17. Hoang, M. L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **113**, 9846–9851 (2016).
18. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
19. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
20. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
21. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
22. Tempest, N., Maclean, A. & Hapangama, D. K. Endometrial stem cell markers: current concepts and unresolved questions. *Int. J. Mol. Sci.* **19**, E3240 (2018).
23. CRUK. Uterine cancer risk. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer/risk-factors#heading-Zero (accessed 28 March 2020).
24. Morice, P., Leary, A., Creutzberg, C., Abu-Rustum, N. & Darai, E. Endometrial cancer. *Lancet* **387**, 1094–1108 (2016).
25. Le Gallo, M. & Bell, D. W. The emerging genomic landscape of endometrial cancer. *Clin. Chem.* **60**, 98–110 (2014).
26. Onstad, M. A., Schmandt, R. E. & Lu, K. H. Addressing the role of obesity in endometrial cancer risk, prevention, and treatment. *J. Clin. Oncol.* **34**, 4225–4230 (2016).
27. Setiawan, V. W. et al. Type I and II endometrial cancers: have they different risk factors? *J. Clin. Oncol.* **31**, 2607–2618 (2013).
28. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
29. The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
30. Anglesio, M. S. et al. Cancer-associated mutations in endometriosis without cancer. *N. Engl. J. Med.* **376**, 1835–1848 (2017).
31. Lac, V. et al. Iatrogenic endometriosis harbors somatic cancer-driver mutations. *Hum. Reprod.* **34**, 69–78 (2019).
32. Lac, V. et al. Oncogenic mutations in histologically normal endometrium: the new normal? *J. Pathol.* **249**, 173–181 (2019).
33. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822–825 (2010).
34. Barker, N. et al. Lgr5^{+ve} stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. *Cell Stem Cell* **6**, 25–36 (2010).
35. Rouhani, F. J. et al. Mutational history of a human cell lineage from somatic to induced pluripotent stem cells. *PLoS Genet.* **12**, e1005932 (2016).
36. Nicholson, A. M. et al. Fixation and spread of somatic mutations in adult human colonic epithelium. *Cell Stem Cell* **22**, 909–918 (2018).
37. Zhang, Y. et al. A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. *Cell Rep.* **24**, 515–527 (2018).
38. Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).
39. Wu, Q. J. et al. Parity and endometrial cancer risk: a meta-analysis of epidemiological studies. *Sci. Rep.* **5**, 14243 (2015).
40. Greaves, M. In utero origins of childhood leukaemia. *Early Hum. Dev.* **81**, 123–129 (2005).
41. Greaves, M. Pre-natal origins of childhood leukemia. *Rev. Clin. Exp. Hematol.* **7**, 233–245 (2003).
42. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal. *Cell* **173**, 611–623 (2018).
43. Anderson, N. D. et al. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science* **361**, eaam8419 (2018).
44. Maura, F. et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat. Commun.* **10**, 3835 (2019).
45. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Sample collection

Anonymized snap-frozen endometrial tissue samples were obtained from five cohorts. All samples were collected in line with the protocols approved by the relevant ethics committees. These are outlined below. All endometrial biopsies underwent formal review by two pathologists, which confirmed benign histology.

Cohort 1. Samples from individuals PD37605, PD37601, PD37607, PD37613, PD37594, PD37595, PD41871, PD41860, PD41857, PD41865, PD41868, PD41859, PD41861 and PD41869 (age 29 to 46) were collected using a Wallach Endocell endometrial sampler from women undergoing examination at the Tommy's National Early Miscarriage Centre, University Hospitals Coventry and Warwickshire NHS Trust. Informed consent was obtained and biopsies collected and stored at the Arden Tissue Bank, University Hospitals Coventry and Warwickshire NHS Trust, in line with the protocols approved by the NRES Committee South Central Southampton B (REC reference 12/SC/0526, 19/04/2013).

Cohort 2. Samples from individuals PD40535, PD39444, PD39953, PD39952, PD39954, PD40107, PD42746 and PD42475 (age 24 to 74) were collected from residual tissues from non-uterine transplant organ donors with ethical and informed consent obtained from the donor's family (REC reference 15/EE/0152 NRES Committee East of England – Cambridge South).

Cohort 3. Individuals PD36804 and PD36805 (aged 47 and 49), underwent total abdominal hysterectomy for benign non-endometrial pathologies and biopsies were collected, snap-frozen and stored at the Human Research Tissue Bank, Cambridge University Hospitals NHS Foundation Trust, in line with the protocols approved by the NRES Committee East of England (REC reference 11/EE/0011, 11/03/2011).

Cohorts 4 and 5. Samples from individuals PD37506, PD38812, PD37507 and PD40659 (age 19 to 81) were obtained at autopsy following death from nongynaecological causes. The use of this material was approved by the London, Surrey Research Ethics Committee (REC reference 17/LO/1801, 26/10/2017) and East of Scotland Research Ethics Service (REC reference 17/ES/0102, 27/07/2017).

Laser-capture microdissection of endometrial glands

Frozen and paraffin sections were used for laser-capture microdissection (LCM). For frozen sections, endometrial tissue was embedded in optimal cutting temperature (OCT) compound. Fourteen-to-twenty-micrometre-thick sections were generated at −20 °C to −23 °C, mounted on to poly-ethylene naphtholate (PEN)-membrane slides (Leica), fixed with 70% ethanol, washed twice with phosphate-buffered saline (PBS), and stained with Gill's haematoxylin for 20 s and eosin for 20 s.

For paraffin sections, frozen endometrial tissue was first thawed at 4 °C for 10–15 min, then fixed in 70% ethanol and embedded in paraffin using standard histological tissue processing. Eight-to-ten-micrometre-thick sections were subsequently cut, mounted onto PEN-membrane slides, and stained by sequential immersion in the following: xylene (2 min, twice), ethanol (100%, 1 min, twice), deionized water (1 min, once), Gill's haematoxylin (10–20 s), tap water (20 s, twice), eosin (10 s, once), tap water (10–20 s, once), ethanol (70%, 20 s, twice) and xylene or neo-clear xylene substitute (10–20 s, twice).

Using a laser-capture microscope (Leica LMD7), individual endometrial glands were first visualized, then dissected (power 7, aperture 1, pulse 119 and speed 5) and collected into separate wells in a 96-well plate. Overview pre- and post-dissection images were taken. In addition, 200–500-μm^2 sections of either myometrium, endometrial stroma or Fallopian tube epithelium were also obtained.

Cell lysis, DNA extraction and whole-genome sequencing of endometrial glands

In brief, 20 µl of an in-house-produced lysis buffer containing 30 mM Tris-HCl pH 8.0 (Sigma Aldrich), 0.5% Tween-20 (Sigma Aldrich), 0.5% NP-40/IGEPAL CA-630 (Sigma Aldrich) and 1.25 µg/ml proteinase K (Qiagen) was added to each well, vortexed (30 s) and spun down at 18 °C (1 min at 1,500 rpm). Samples were subsequently incubated in a thermal cycler for 60 min at 50 °C and 30 min at 75 °C before storage at −80 °C.

All samples in this study were processed using a recently developed low-input enzymatic fragmentation-based library preparation method[14,15]. In brief, each 20 µl LCM lysate was mixed with 50 µl Ampure XP beads (Beckman Coulter) and 50 µl TE buffer (Ambion; 10 mM Tris-HCl, 1 mM EDTA) at room temperature. Following a 5-min binding reaction and magnetic bead separation, genomic DNA was washed twice with 75% ethanol. Beads were resuspended in 26 µl TE buffer and the bead–genomic DNA slurry was processed immediately for DNA library construction. Each sample (26 µl) was mixed with 7 µl of 5× Ultra II FS buffer, 2 µl of Ultra II FS enzyme (New England BioLabs) and incubated on a thermal cycler for 12 min at 37 °C, then 30 min at 65 °C. Following DNA fragmentation and A-tailing, each sample was incubated for 20 min at 20 °C with a mixture of 30 µl ligation mix and 1 µl ligation enhancer (New England BioLabs), 0.9 µl nuclease-free water (Ambion) and 0.1 µl duplexed adapters (100 µM; 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3′ (in which * is a phosphorothioated DNA base), 5′-phos-GATC GGAAGAGCGGTTCAGCAGGAATGCCGAG-3′). Adaptor-ligated libraries were purified using Ampure XP beads by addition of 65 µl Ampure XP solution (Beckman Coulter) and 65 µl TE buffer (Ambion). Following elution and bead separation, DNA libraries (21.5 µl) were amplified by PCR by addition of 25 µl KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 1 µl PE1.0 primer (100 µM; 5′-AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATC*T-3′) and 2.5 µl iPC R-Tag (40 µM; 5′-CAAGCAGAAGACGGCATACGAGATXGAGATCGGTCTC GGCATTCCTGCTGAACCGCTCTTCCGATC-3′) in which 'X' represents one of 96 unique 8-base indexes The samples were then mixed and thermal-cycled as follows: 98 °C for 5 min, then 12 cycles of 98 °C for 30 s, 65 °C for 30 s, 72 °C for 1 min and finally 72 °C for 5 min. Amplified libraries were purified using a 0.7:1 volumetric ratio of Ampure Beads (Beckman Coulter) to PCR product and eluted into 25 µl of nuclease-free water (Ambion). DNA libraries were adjusted to 2.4 nM and sequenced on the HiSeq X platform (illumina) according to the manufacturer's instructions, with the exception that we used iPCRtag-seq (5′-AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC-3′) to read the library index.

All LCM samples were subjected to whole genome sequencing of 15–40.3×, using 150-base-pair clipped reads sequenced on HiSeq X platform (Illumina). For selected donors (PD36804, PD36805 and PD37506), bulk samples (fragments of uterus or cervix) were also whole-genome sequenced.

Variant calling

Substitutions. Sequencing data were aligned to the reference human genome (NCBI build 37) using the Burrow–Wheeler Aligner (BWA-MEM)[46]. Duplicates were marked and removed, and mapping quality thresholds were set at 30. Single-base somatic substitutions were called using 'Cancer Variants through Expectation Maximization' (CaVEMan) algorithm (major copy number 5, minor copy number 2)[47]. To exclude germline variants, matched normal samples (cervix, myometrium, Fallopian tube or endometrial stroma) were used to run the algorithm.

A set of previously described post-processing filters were subsequently applied: (1) to remove common single nucleotide polymorphisms, variants were filtered against a panel of 75 unmatched normal samples[47]; (2) to remove mapping artefacts associated with BWA-MEM, the median alignment score of reads that support a mutation should be greater than or equal to 140 (ASMD ≥ 140) and fewer than half of the reads should be clipped (CLPM = 0)[14]; (3) to remove artefacts that are specific to the library preparation for LCM samples, two additional filters were used. A fragment-based filter, which is designed to remove overlapping reads that result from the relatively shorter insert sizes allowed in this protocol that can lead to double counting of variants, and a cruciform filter, which removes erroneous variants that can be introduced due to the incorrect processing of cruciform DNA. For each variant, the standard deviation (s.d.) and median absolute deviation (MAD) of the variant position within the read was calculated separately for positive and negative strands reads. If a variant was supported by a low number of reads for one strand, the filtering was based on the statistics calculated from the reads derived from the other strand and it was required that either: (a) ≤90% of supporting reads report the variant within the first 15% of the read as determined from the alignment start, or (b) that the MAD > 0 and s.d. > 4. Where both strands were supported by sufficient reads, it was required for both strands separately to either: (a) ≤90% of supporting reads report the variant within the first 15% of the read as determined from the alignment start, (b) that the MAD > 2 and s.d. > 2, or (c) that at least one strand has fulfilled the criteria MAD > 1 and s.d. > 10.

Validation experiments and sensitivity. To validate somatic variants, for selected donors, pairs of biological 'near-replicates' were obtained. For these experiments, we collected two samples from the same endometrial gland that was identified on two or more consecutive levels using a *z*-stacking approach; each sample was processed separately with independent DNA extraction, library preparation and whole-genome sequencing. As these samples were obtained from the same glands, they would represent derivatives of the same clone and therefore the same sensitivity would be assumed in both samples in each pair. The maximum likelihood estimate for sensitivity (*S*) was then calculated as follows:

$$S = \frac{2n_2}{n_1 + 2n_2}$$

in which n_1 is the number of variants called only in one of the two LCM samples and n_2 is the number of variants called in both LCM samples in each pair. Using this approach, the mean sensitivity of somatic mutation variant calling was estimated at >86% (range 0.70–0.95%).

Indels. Insertions and deletions were called using cgpPindel[48,49]. To remove germline variants, the algorithm was run with the same matched normal samples that were used for calling substitutions. Post-processing filters were applied as previously described[47]. In addition, a 'Qual' filter (the sum of the mapping qualities of the supporting reads) of at least 300 and a depth cut-off of at least 15 reads were used.

Copy-number variants and structural variants. Allele-specific copy-number profiles were reconstructed for the endometrial gland samples by ASCAT[50,51] using matched samples as described in 'Substitutions', ploidy of 2 and contamination with other cell types of 10%. Only samples with a minimum coverage of 15× and above were used. All putative copy-number changes were visually inspected for copy-number profiles on Jbrowse[52].

Structural variants in endometrial glands were called using matched samples (as described in 'Substitutions') with the 'Breakpoints Via Assembly' (BRASS) algorithm, and further annotated by GRASS (https://github.com/cancerit/BRASS). Potential structural variants are detected for the sample of interest and read-pair clusters that support the structural variant are used for breakpoint-sequence de novo assembly. Absence of supporting evidence in the matched control indicates that the structural variant was acquired in the sample of interest. The isolation of minute amounts of DNA for sequencing in combination with the LCM enzymatic-fragmentation-based library preparation procedure introduces additional artefacts and additonal post-processing filtering was performed in two phases:

Further annotation of structural variants with statistics that detect LCM-specific artefacts. All structural variants detected by BRASS were further annotated with AnnotateBRASS (https://github.com/MathijsSanders/AnnotateBRASS). Each structural variant is defined by two breakpoints and their genomic coordinates.

First, the following statistics were determined for each breakpoint separately: (1) the total number of reads supporting the structural variant; (2) the total number of unique reads supporting the structural variant, based on alignment position and read orientation; (3) the standard deviation of the alignment positions of reads supporting the structural variant; (4) the number of chromosomes, based on read-pairs not supporting the structural variant, to which one read mapped while the mate-read aligned to the structural variant breakpoint; (5) the number of reads supporting the structural variant that had an alternative alignment (XA-tag); (6) the number of reads supporting the structural variant that had an alternative alignment score (XS-tag) similar to the current alignment score; (7) the percentage of read-pairs not supporting the structural variant with a discordant inferred insert size (default: ≥1,000 bp).

Second, a wider search for read-pairs supporting the structural variant was intiated and the following statistics were calculated for each breakpoint separately: (1) the total number of reads supporting the structural variant; (2) the total number of unique reads supporting the structural variant, based on alignment position and read orientation; (3) the standard deviation of the alignment positions of reads supporting the structural variant; (4) the number of reads supporting the structural variant that had an alternative alignment; (5) the number of reads supporting the structural variant that had an alternative alignment score similar to the current alignment score. Third, reads spanning the structural variant breakpoints are often clipped. Clipped sequences of sufficient length can be aligned to other positions on the genome (that is, supplementary alignment) and it is expected that these align to the proximity of the other structural variant breakpoint. Based on the clipping positions and supplementary alignments, the following was determined for each structural variant: (1) whether the clipped sequences of read-pairs spanning a structural variant breakpoint align in the proximity of the other structural variant breakpoint; (2) whether the clipping within read-pairs supporting the structural variant occurred at roughly the same genomic position (default: all clipping positions occurred within 10 bp of each other).

Next, BRASS uses a single matched control and a panel of normal samples (bulk whole-genome sequencing) to determine whether a structural variant is somatic. Structural variants observed in the sample of interest but not in the matched control or panel of normal samples are considered somatic. However, owing to the difference in library preparation and the variance of spatial genomic coverage observed, it is not always possible to accurately assess the validity of the structural variant. Two approaches were implemented to determine whether the structural variant is somatic: (1) a wider search in the matched control sample was performed to search for read-pairs that could support the structural variant. The structural variant was still considered detected in cases in which the discovered read-pairs were insufficient for breakpoint sequence de novo assembly. (2) Additional controls can be defined in cases for which multiple samples have been isolated for the same individual. Samples from the same individual with little genetic relationship, as determined from the single-nucleotide variants (SNVs) and indels, can be used as controls to determine whether the detected structural variant is germline or a recurrent artefact.

Post hoc filtering of structural variants based on a combination of the above statistics. Structural variants were further filtered on the basis of the statistics described in 'Further annotation of structural variants with statistics that detect LCM-specific artefacts'. The optimal set of statistics and their most practical thresholds depend on the achieved coverage and stringency of filtering desired. As a default, the following criteria were used for detecting somatic structural variants: (1) for each breakpoint there must be ≥4 unique reads supporting the structural variant; (2) the alignment position standard deviation must be >0; (3) at each breakpoint there are read-pairs not supporting the structural variant that map to <5 other chromosomes (4); (5) the total number of chromosomes mapped to by read-pairs not supporting the structural variant for both breakpoints should be <7; (6) the percentage of reads supporting the structural variant with alternative alignments or alternative alignments with similar alignment scores should be ≤50% for both structural variant breakpoints separately; (7) the percentage of discordant read-pairs not supporting the structural variant should be ≤7.5% of total read-pairs for both structural variant breakpoints separately; (8) for the wider search of structural-variant-supporting read-pairs the same thresholds apply as under criteria 1–6; (9) there are no read-pairs in the matched control that support the structural variant; (10) the structural variant is not detected in any of the other control samples, or there were ≤2 samples carrying the same structural variant and the proportion of control samples carrying the structural variant was <1/3 of the defined control set; (11) read-pairs supporting the structural variant were not allowed to have widely divergent clipping positions in terms of genomic location for both structural variant breakpoints separately.

Formal clonality assessment of individual endometrial glands

We applied the DPClust v.2.2.7[47] subclonal reconstruction caller with default parameters to the SNVs in each endometrial crypt to assess the clonality of each crypt. SNVs that fell within a detected copy-number alteration were excluded from this analysis. The purity of each crypt was set to 1, the resulting mutation clusters therefore represent proportions of the overall sequenced cells. For every sample, this analysis yields the number of mutation clusters and assigned mutations, and the proportion of overall cells that each cluster represents.

A gland was determined to be the result of a single progenitor cell if a single mutation cluster was obtained or when the proportions of cells in which multiple mutation clusters were detected dictate a linear relationship. Akin to the so-called 'pigeonhole principle'[53], in such a scenario the sum of the estimated proportions of cells of a pair of cellular populations exceeds 1 (100% of cells), which means at least some cells must contain the mutations in both clusters. Alternatively, if the sum of the estimated proportions does not exceed 1 the populations could be the result of a single or of separate ancestors.

The analysis showed that 89.9% (231 out of 257) of samples had a major clone (defined as those with ≥75% of sequenced cells) with clusters containing on average 79.5% of all substitutions (s.d. = 24.9%) (Extended Data Fig. 2, Supplementary Results 3). Eighty-three per cent (214 out of 257) of glands showed evidence of a further, subclonal cell population which—based on the pigeonhole principle—is a descendant of the main clonal population. The majority of glands also showed minor contamination by cells that do not share somatic mutations with the observed clonal expansions, potentially including endometrial stromal cells, inflammatory cells and epithelial cells from other glands.

Detection of driver mutations

Analysis of driver variants in the normal endometrial glands was performed in two parts. First, filtered CaVEMan and Pindel variants were intersected against a previously published list of 369 genes that are under selection in human cancers[28]. All nonsynonymous mutations were annotated to indicate mode of action using a Cancer Gene Census (719 genes) and a catalogue of 764 genes (https://www.cancergenomeinterpreter.org). Truncating variants (nonsense, frameshift and essential splice), which resided in recessive or tumour-suppressor genes were declared likely drivers. Missense mutations in recessive or tumour-suppressor genes and dominant genes or oncogenes were triaged against a database of validated hotspot mutations (http://www.cbioportal.org/mutation_mapper). All mutations that were shown to be known mutational hotspots or 'probably oncogenic' were declared drivers (Supplementary Results 15). In addition, identified activating mutations in mutational hotspots in the gene *RRAS2*, which involves the RAS/MAPK pathway, were declared as likely drivers.

Second, to identify genes that are under positive selection in normal endometrium, we used the dN/dS[28] method, which is based on the observed:expected ratios of nonsynonymous to synonymous mutations. The analysis was carried out for the whole genome ($q < 0.01$ and $q < 0.001$) and for 369 known cancer genes[28] (restricted hypothesis testing, $q < 0.05$). Twelve genes were found to be under positive selection in normal endometrial glands. The output of this analysis was also used to assess whether missense mutations in genes that are under positive selection in normal and/or malignant endometrium (*PIK3CA*, *ERBB2*, *ERBB3*, *FBXW7* and *CHD4*) but are not known mutational hotspots, are likely to be drivers. We calculated the fraction of the mutations tested that are likely to be drivers (f) using the following equation: $f = (w - 1)/w$, in which w is the observed missense count (52) divided by the expected count (0.14). If f was ≥ 0.95, then all missense mutations in that gene were declared likely drivers. To compare patterns of selection in normal endometrial epithelium and cancer, we performed dN/dS analysis on previously published data from the The Cancer Genome Atlas (TCGA)[29].

Phylogenetic tree reconstruction

Phylogenies for endometrial glands were reconstructed for 17 donors. Owing to the low number of available samples, donor PD38812 was not included in this analysis. We first generated trees using substitutions called by CaVEMan; matched normal samples were used to exclude germline variants and post-processing filters were applied as in 'Substitutions'. Final variants were recalled in all samples from each donor using an in-house re-genotyping algorithm (cgpVAF). Variants with a VAF >0.3 were noted to be present ('1'), VAF <0.1 absent ('0') and between 0.1 and 0.3 as ambiguous ('?'). This approach excludes private subclonal variants from the tree building. The tree was reconstructed using a maximum parsimony approach[54] and branch support was calculated using 1,000 bootstrap replicates. Nodes with a confidence lower than 50 were collapsed into polytomies and branch lengths of the collapsed tree were determined by the number of assigned substitutions.

The constructed phylogenies were validated using indels called by Pindel and filtered as in 'Indels'. The same approach was applied for the final indel matrices. Although the lower number of indels resulted in more polytomous tree, the overall tree topologies were reconcilable with those generated using substitutions.

Cancer driver mutations, and copy-number and structural variants were annotated manually in the trees.

Timing of driver mutations

To estimate the time interval in which specific driver mutations occurred, we applied two approaches: (a) 'patient-based', in which we calculated a patient-specific mutation rate by taking the ratio of the mean mutation burden per endometrial gland of the patient and the age of the patient; (b) 'cohort-based', in which the mutation rate for each patient was derived from the linear mixed-effect model for total mutation rate that included data from the entire cohort (Supplementary Results 5). The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. Both approaches rely on the assumption of a constant mutation rate for endometrial glands throughout the life of the patient.

Mutational signature analysis

Mutational signature extraction was performed using mutations assigned to every branch of the reconstructed phylogenetic trees and each branch was treated as an individual sample. Such an approach allows the characterization and differentiation of specific mutational processes that were operative at various times in individual glands. Substitutions were first categorized into 96 classes following the method used by the 'Mutational Signatures' working group of PCAWG[4]. SBS signature analysis was performed in three steps: extraction, deconvolution and re-attribution. SBS signatures were extracted using three approaches: (1) using the HDP package (https://github.com/nicolaroberts/hdp) that uses the hierarchical Bayesian Dirichlet process either de novo, or (2) with reference signatures ('priors') identified by the Mutational Signatures working group of PCAWG[4]; and (3) nonnegative matrix factorization[4]. (1) HDP de novo signature extraction revealed theree components (components 1, 2 and 0 in Extended Data Fig. 4a). The similarity of the components to the 65 reference signatures was assessed; component 2 had a high cosine similarity (>0.95) to SBS18. (2) HDP signature extraction with all 65 PCAWG priors yielded the following components: 'priors' or reference SBS signatures (P1 = SBS1, P5 = SBS5, P18 = SBS18, P23 = SBS23 and P40 = SBS40); a 'new' component that did not match any of the provided 65 reference signatures or priors (N1) and 'component 0' (comp 0); all of the components from this extraction were taken for further analysis and deconvolution (Extended Data Fig. 4, b). Because P1, P5, P18, P23 and P40 showed high cosine similarity (>0.95) to the respective signatures (SBS1, SBS5, SBS18, SBS23 and SBS40), no further deconvolution of these components was required. As component N1 did not show high cosine similarity to any of the reference signatures, deconvolution was performed using a 'deconvolution' catalogue comprising all of the extracted signatures (SBS1, SBS5, SBS18, SBS23 and SBS40). Final exposures were derived and signatures re-attributed to the individual samples (branches). As SBS5 and SBS40 are relatively featureless and present particular challenges in estimating their separate contributions (as previously outlined[4]), these have therefore been combined (but are shown separately in Supplementary Results 9). SBS23 was previously found in a small number of liver cancers at high mutation burdens. Given its low mutation burden and small contribution here, it is unclear whether this is really the same signature and were therefore included it in the unattributable category. (3) Nonnegative matrix factorization signature extraction was performed using SigprofilerExtractor Version 0.0.5.51 (https://pypi.org/project/sigproextractor/#history), SigprofilerMatrixGenerator Version 1.0.2 (https://pypi.org/project/SigProfilerMatrixGenerator/#history) and SigprofilerPlotting Version 1.0.3 (https://pypi.org/project/sigProfilerPlotting/) on solutions between 1 and 20 signatures with 3 signatures chosen as the optimal solution running 1,000 iterations. The extraction yielded 3 signatures, which were further deconvoluted as following: signature A into SBS1 (8.16%), SBS5 (79.88%) and SBS23 (11.96%); signature B into SBS1 (16.18%), SBS5 (22.6%) and SBS18 (61.22%); and signature C into SBS1(42.1%) and SBS5(57.9%).

Indels were classified using the PCAWG method[4] and composite mutational spectra were generated for each donor. However, given the relatively low numbers of indels, no formal signature extraction was performed.

Calculations of mutation burden and estimation of mutation rate

For these analyses, we excluded the following cases: (a) samples from donors with missing metadata (body mass index and parity); (b) samples with an adjusted coverage (variant allele fraction depth) of <7.5 (adjusted coverage defined as VAF × sequencing depth). To account for the nonindependent sampling per patient, we used mixed effects models. In these analyses, we tested features either with a known effect on mutation burden or endometrial cancer risk; age, read depth and VAF, body mass index, and parity. In addition, we tested whether there was any significant difference between different patient cohorts. Finally, we tested whether menstrual phase has an effect on the clonality and mutation burdens. All statistical analyses were performed in R and are summarized in the Supplementary Results.

To compare somatic mutations of normal endometrium to that of endometrial cancer, we used previously described variants from PCAWG[4].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Whole-genome sequencing data are deposited in the European Genome–Phenome Archive (EGA) with accession number EGAS00001002471.

Code availability

Code for statistical analyses on total substitution and driver mutation burdens is included in the Supplementary Information, and is deposited on GitHub at https://github.com/LuizaMoore/Endometrium. All other code is available on request.

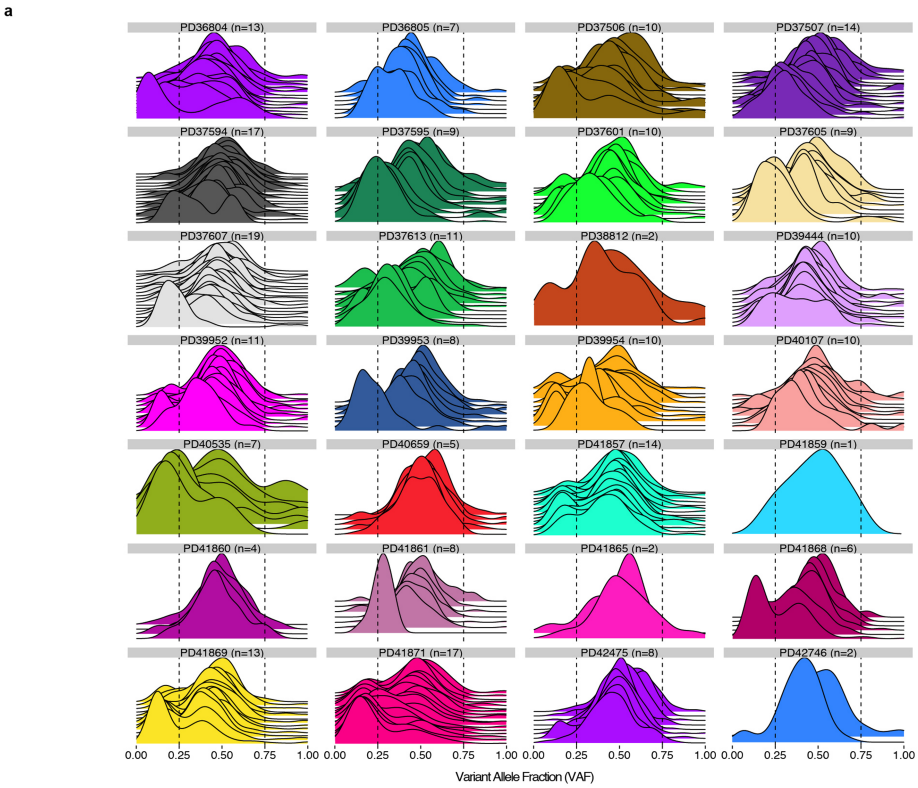
46. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
47. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
48. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.2 (2015).
49. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
50. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
51. Raine, K. M. et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinformatics* **56**, 15.19.11–15.19.17 (2016).
52. Buels, R. et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
53. Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
54. Hoang, D. T. et al. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018).

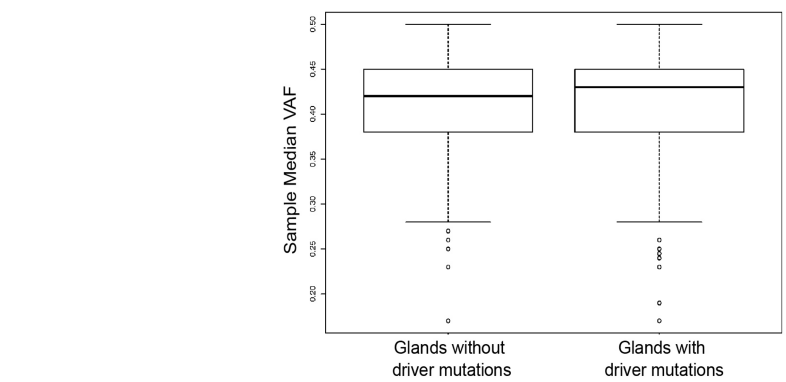
Acknowledgements We thank the staff of WTSI Sample Logistics, Genotyping, Pulldown, Sequencing and Informatics facilities for their contribution; L. O'Neil, C. Latimer and P. Scott for technical support; F. Nadeu, N. Roberts and J. Wang for their advice on mutational signature extraction; A. R. J. Lawson, F. Abascal and S. Grossmann for their assistance with data analysis; and the Cambridge Biorepository for Translational Medicine for the provision of samples from deceased transplant organ donors. This work was supported by the Wellcome Trust. L.M. is a recipient of a CRUK Clinical PhD fellowship (C20/A20917) and Pathological Society of Great Britain and Ireland Trainee Small Grant (grant reference no. 1175). S.F.B. was supported by the Swiss National Science Foundation (P2SKP3-171753 and P400PB-180790). M.A.S. is supported by a Rubicon fellowship from NWO (019.153LW.038).

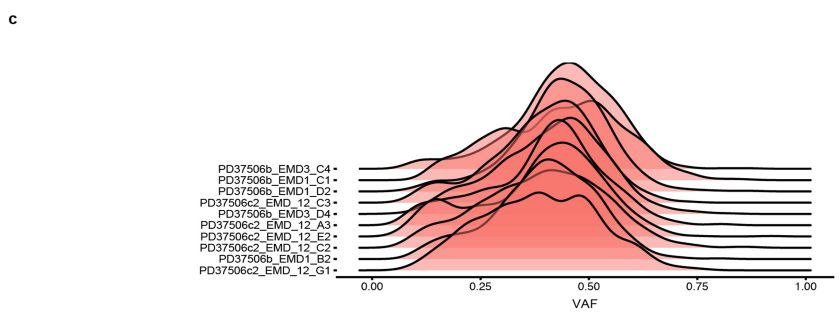
Author contributions M.R.S. and L.M. designed the study and wrote the manuscript with contributions from all authors. K.S.-P., C.A.I.-D., J.J.B., K.T.M., M.J.-L. and L.M. obtained samples. P.E. and L.M. devised the protocol for laser-capture microscopy, DNA extraction and sequencing of endometrial glands. L.M. prepared sections, reviewed histology, and microdissected and lysed endometrial glands. Y.H. assisted with tissue processing and section preparation. L.M. performed data curation and analysis with help from D.L., T.H.H.C., M.A.S., S.C.D., K.J.D., T.B., R.R., T.J.M., J.N., P.S.T., S.F.B. and H.L.-S. T.H.H.C. reconstructed phylogenetic trees. S.C.D. performed formal clonality assessment with dpclust. M.A.S. devised filters for substitutions and structural variants. D.L., F.M. and S.M. assisted with signature analyses. I.M. assisted with statistical and dN/dS analyses. P.J.C. oversaw statistical analyses and performed analysis of structural variants. M.R.S. supervised the study.

Competing interests M.R.S. is on the Scientific Advisory Board for GRAIL.

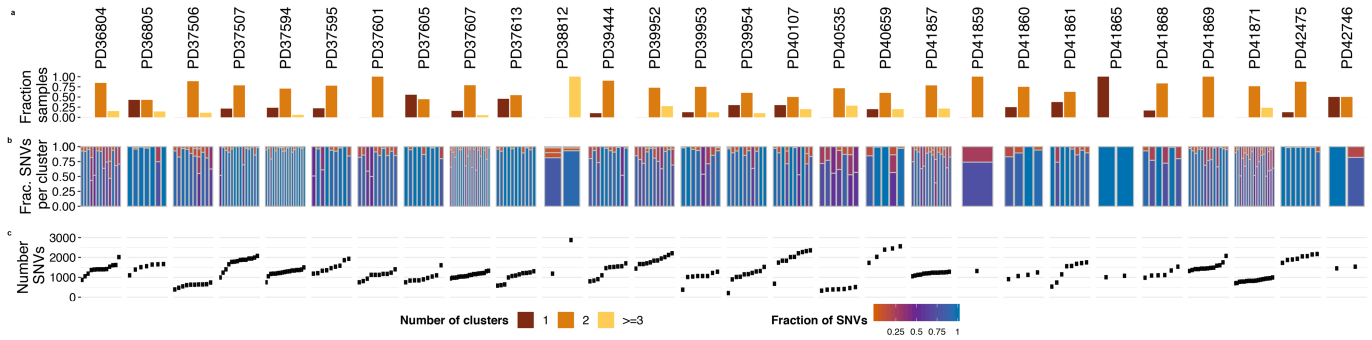
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2214-z.
Correspondence and requests for materials should be addressed to M.R.S.
Peer review information *Nature* thanks Michael Lawrence and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



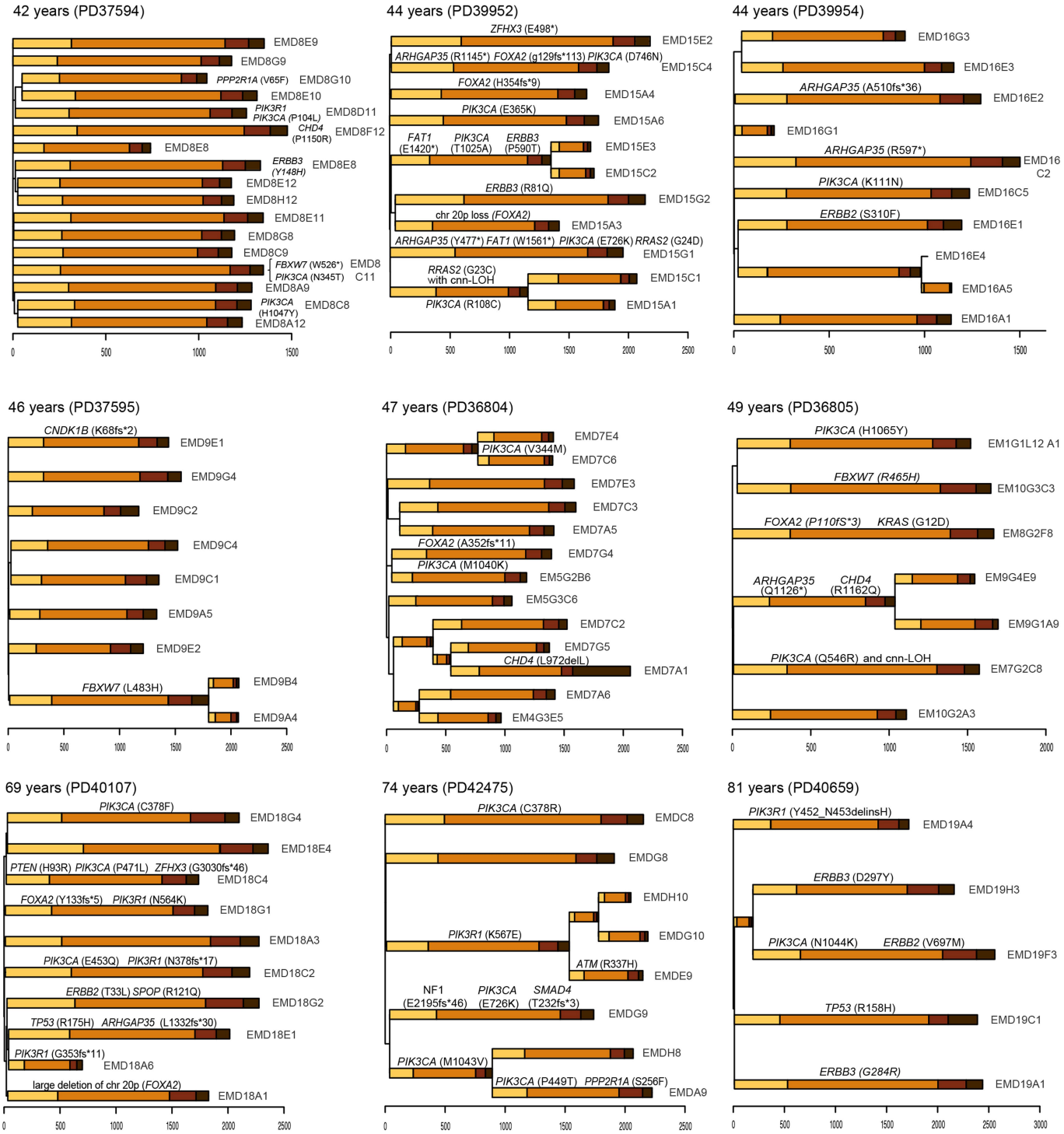




Extended Data Fig. 1 | Clonality of endometrial glands and driver mutations. **a**, The majority of the sampled normal endometrial glands ($n = 257$ individual endometrial glands) were clonal with a median VAF for all identified indels of 0.3 or above. **b**, The presence of a driver mutation did not have a significant effect on the observed monoclonality of the glands (two-sided Mann–Whitney *U*-test, $P = 0.1$). Here, we compared the median VAF of endometrial glands with drivers (median = 0.33, range 0.17–0.5, $n = 145$) to that of glands without drivers (median = 0.31, range 0.16–0.5, $n = 112$). Box plots were constructed with the upper and lower edge of the box defining the 25th (Q1) and 75th (Q3) percentile, respectively, outliers (plotted as circles) are defined as values beyond the whiskers (upper, Q1 − 1.5 × interquartile range (IQR) and lower Q3 + 1.5 × IQR). **c**, All glands from the 19-year-old donor (donor PD37506) ($n = 10$ individual endometrial glands) were clonal with a median VAF ≥ 0.3, but there were no detectable driver mutations.

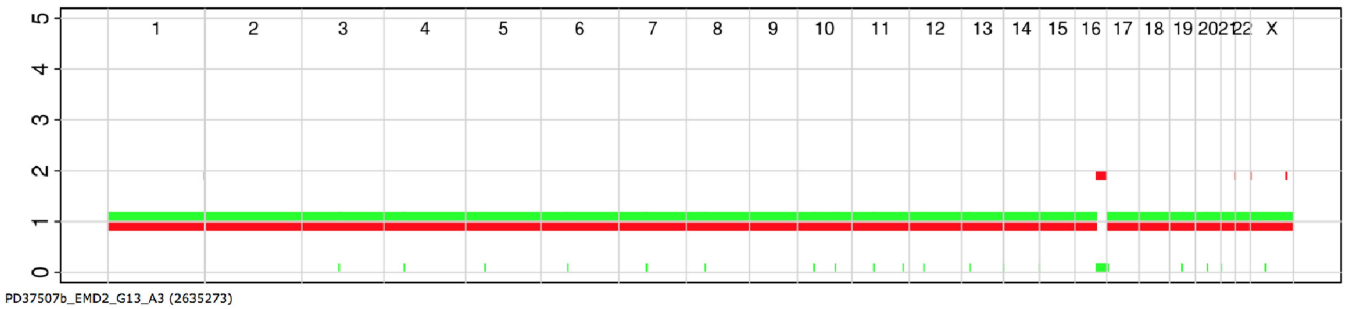


Extended Data Fig. 2 | Assessment of clonal composition of individual endometrial glands using the mutation clustering method dpclust. **a**–**c**, Each column contains a summary of the clonality analysis for individual donors, showing the fraction of samples in which 1, 2 or 3 or more mutation clusters were found (**a**), the fraction of mutations assigned per cluster for each sample (**b**) and at the total number of SNVs per sample (**c**).

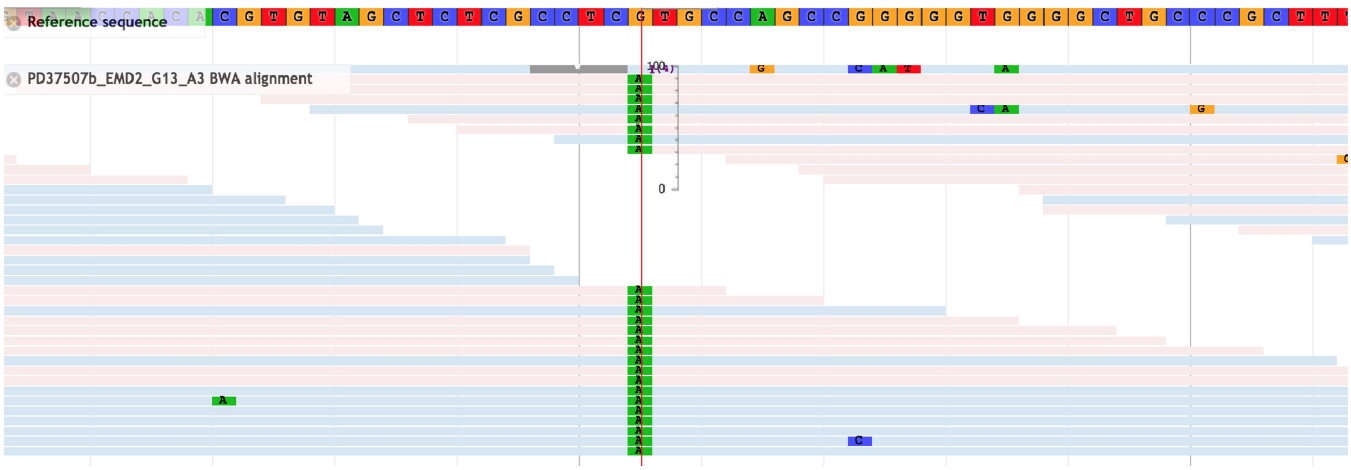


Extended Data Fig. 3 | Phylogenetic trees of endometrial glands for donors aged 42 to 81 years. Phylogenetic trees for twelve donors aged 42 to 81 years were also reconstructed using SBSs with branch length proportional to the number of variants; the stacked bar plots represent the attributed SBS mutational signatures that contributed to each branch. Signature extraction was not performed on branches with fewer than 100 substitutions. The ordering of signatures within each branch is for visualization purposes only. Every single studied gland from donors PD39952 (44 year old) and PD40659 (81 year old) had at least one driver mutation.

a

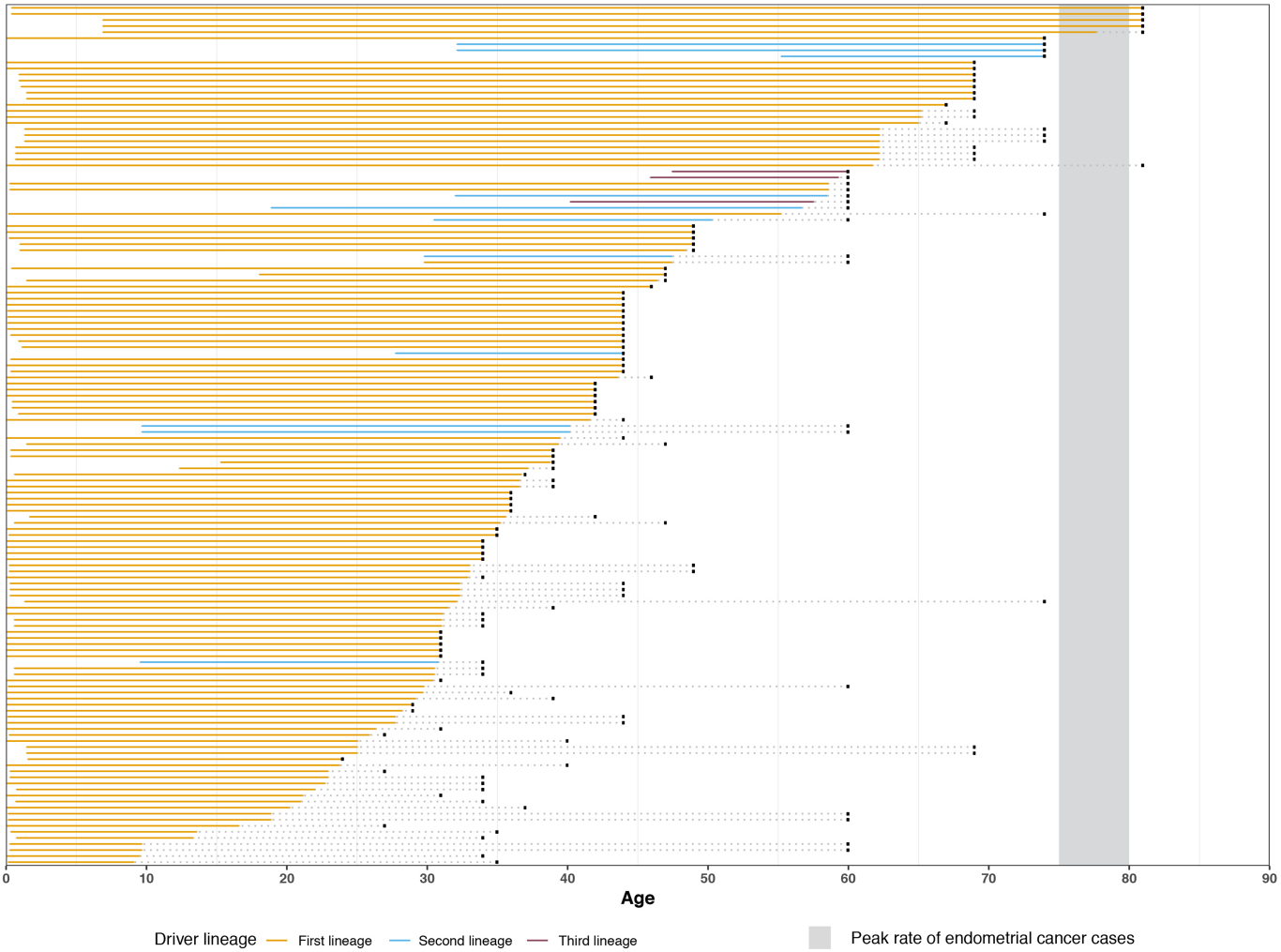


b

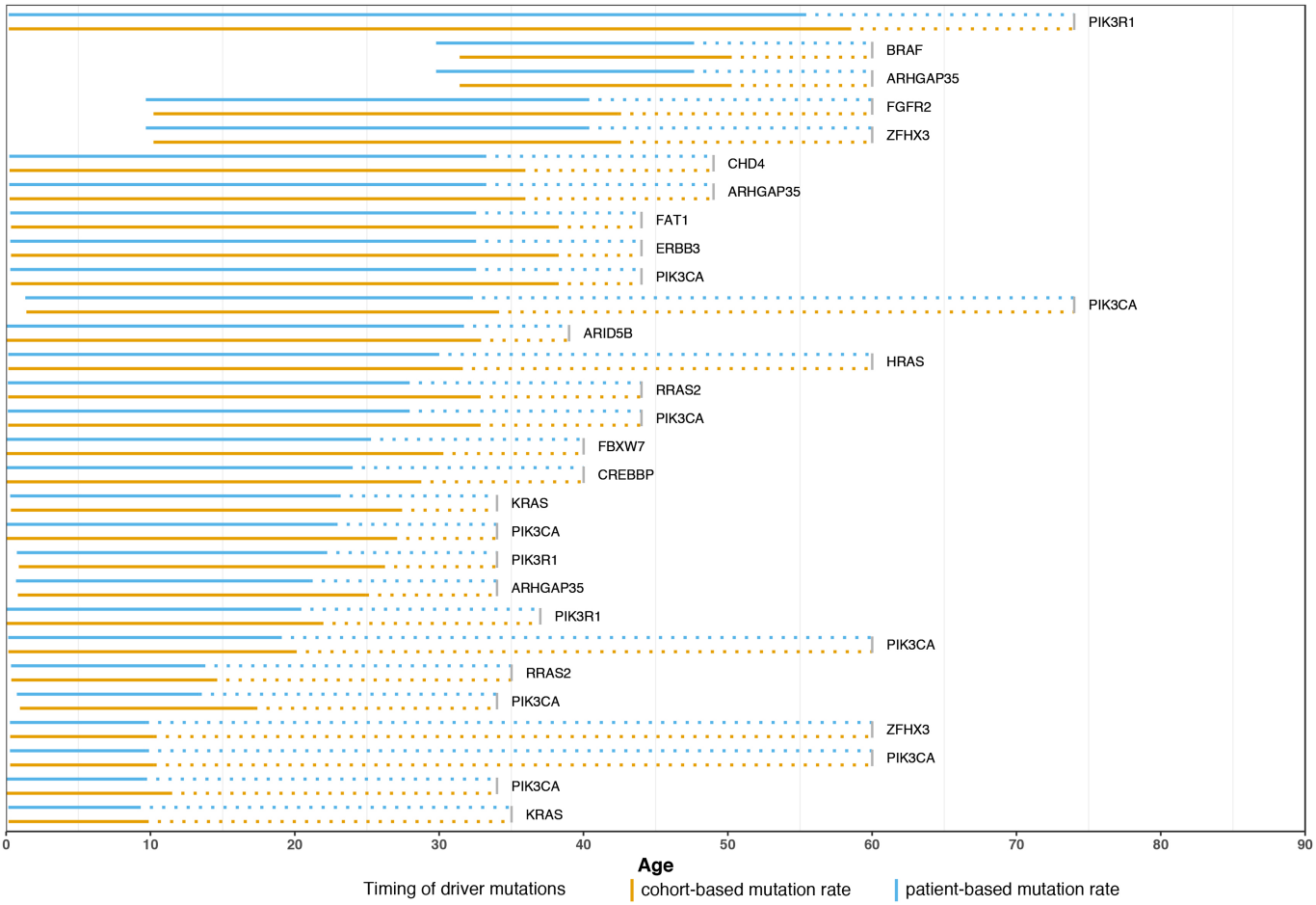


Extended Data Fig. 4 | An example of copy-number neutral loss of heterozygosity in a normal endometrial gland. a, A biallelic truncating mutation is seen in *ZFHX3* (p.R715*), with every read carrying the variant. **b**, Associated copy-number neutral loss of heterozygosity is observed on chromosome 16.

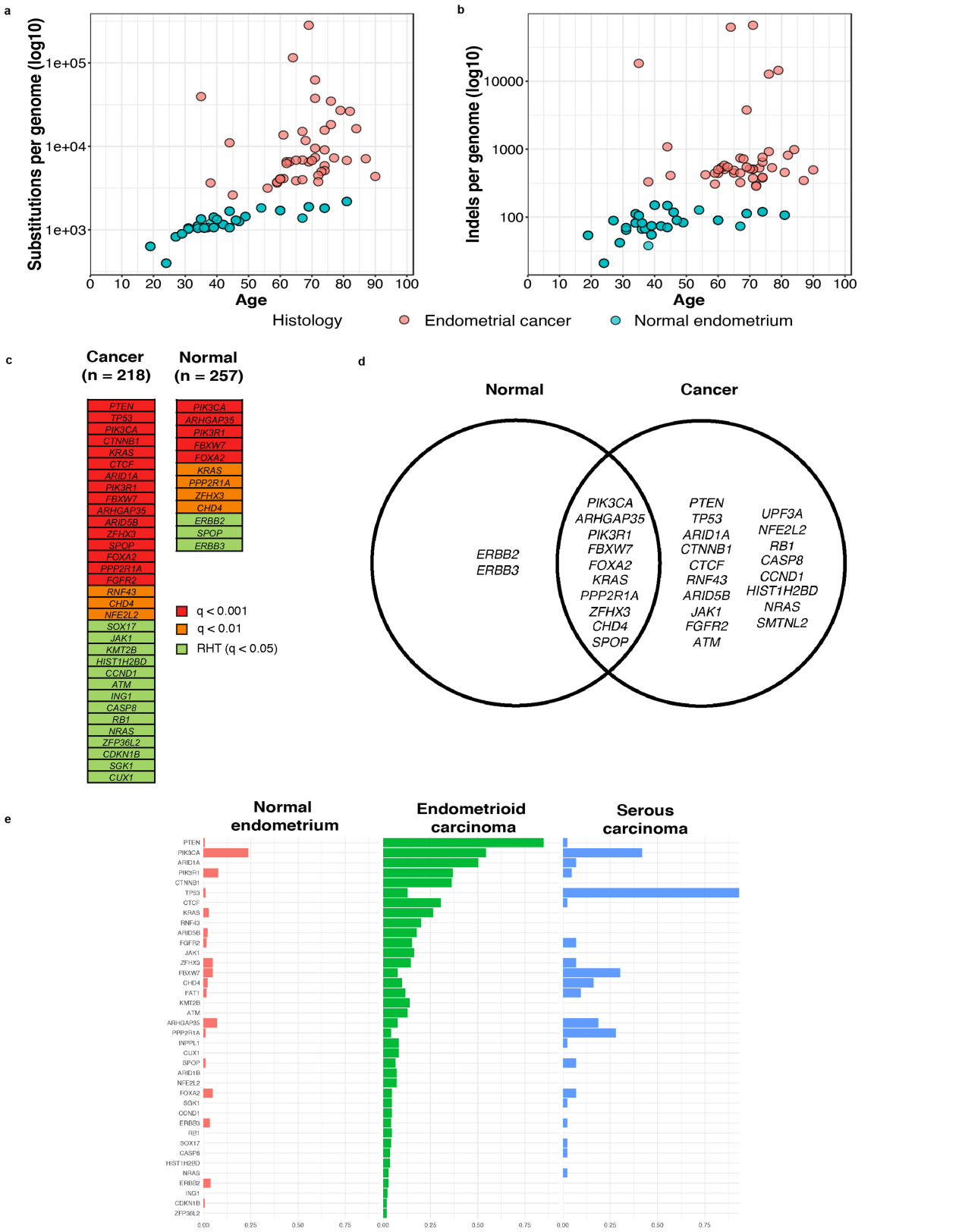
Extended Data Fig. 5 | Oncoplot of all driver mutations and their distribution across individual endometrial-gland samples and donors. Each cell represents an individual endometrial gland sample and is colour-coded to represent the total number of detected driver mutations (0–3). *PIK3CA* was the most frequently mutated gene, with at least 1 mutation detected in 54% (15 out of 28) of women. In some glands, these mutations in *PIK3CA* co-occurred with mutations in *ZFHX3*, *ARHGAP35*, *FGFR2*, *FOXA2* and other genes that are also selected for in endometrial cancer.



Extended Data Fig. 6 | Timing of all driver mutations. To time the driver mutations, we used the reconstructed SNV-based phylogenetic trees for 25 out the 28 donors. To estimate the time interval in which specific mutations occurred, we calculated a patient-specific mutation rate by taking the ratio of the mean mutation burden per endometrial gland of the patient and the age of the patient. The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. This approach relies on the assumption of a constant mutation rate for endometrial glands throughout the life of a patient. The same approach was used for dating indels. We dated the driver mutations that occurred in the trunks and branches.



Extended Data Fig. 7 | Timing of driver mutations using patient-based and cohort-based estimates of mutation rates. To estimate the time interval in which specific mutations occurred, we applied two approaches: (a) 'patient-based', in which we calculated a patient-specific mutation rate by taking the ratio of the mean mutation burden per endometrial gland of the patient and the age of the patient; (b) 'cohort-based', in which the mutation rate for each patient was derived from the linear mixed-effect model for total mutation rate that included data from the entire cohort. The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. Both approaches rely on the assumption of a constant mutation rate for endometrial glands throughout the life of the patient. The dotted line represents time from the upper bound of the estimate to sampling.



Extended Data Fig. 8 | Comparison between normal endometrial epithelium and endometrial cancer. a, **b**, Normal endometrial glands show a lower total mutation burden (substitutions (**a**) and indels (**b**)) than do endometrial cancers (data from PCAWG[2]). **c**, Genes under significant positive selection (dN/dS > 1) in normal endometrial epithelium ($n = 257$ individual endometrial glands) and endometrial cancer ($n = 218$ biologically independent samples). q values were calculated for the whole exome, and under a restricted hypothesis test of 369 known endometrial cancer genes[28]. **d**, Venn diagram showing the overlap between the genes under positive selection in normal endometrium and endometrial cancer from the results described in **c**. **e**, Identified driver mutations and their distribution in normal endometrial glands and the two major types of endometrial cancer (endometroid and serous carcinoma).

nature research

Corresponding author(s): Michael R. Stratton

Last updated by author(s): 08/03/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Image processing from sequencing data using the proprietary Illumina X10 software that is maintained, installed and distributed by Illumina with their X10 platform.
Data analysis	Alignment and variant calling performed using Sanger Institute's custom pipeline. Single-nucleotide substitutions were called using the CaVEMan (cancer variants through expectation maximization) algorithm (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm (https://github.com/genome/pindel). Rearrangements were called using the BRASS (breakpoint via assembly) algorithm (https://github.com/cancerit/BRASS). Clonality assessment was performed using previously described dp clust algorithm (Nik-Zainal, et al, Cell, 2012) . Data analysis was performed in R (3.4.1) with RStudio (1.0.153) Code for statistical analyses on total substitution and driver mutation burdens is included in the supplementary material and is available on GitHub https://github.com/LuizaMoore/Endometrium. Additional code is available on request from the authors. Open source packages used in the data analysis include the following: package * version date source ade4 1.7-11 2018-04-05 CRAN (R 3.4.4) AnnotationDbi 1.38.2 2017-07-27 Bioconductor assertthat 0.2.0 2017-04-11 CRAN (R 3.4.0) backports 1.1.2 2017-12-13 CRAN (R 3.4.3) base * 3.4.1 2017-07-07 local bayesplot 1.6.0 2018-08-02 CRAN (R 3.4.4) bibtex 0.4.2 2017-06-30 CRAN (R 3.4.1) bindr 0.1.1 2018-03-13 CRAN (R 3.4.4) bindrcpp * 0.2.2 2018-03-29 CRAN (R 3.4.4) Biobase * 2.36.2 2017-05-04 Bioconductor

BiocGenerics * 0.22.1 2017-10-07 Bioconductor
BiocInstaller 1.26.1 2017-09-01 Bioconductor
BiocParallel 1.10.1 2017-05-03 Bioconductor
biomaRt 2.32.1 2017-06-09 Bioconductor
Biostrings * 2.44.2 2017-07-21 Bioconductor
bit 1.1-14 2018-05-29 CRAN (R 3.4.4)
bit64 0.9-7 2017-05-08 CRAN (R 3.4.0)
bitops 1.0-6 2013-08-17 CRAN (R 3.4.0)
blob 1.1.1 2018-03-25 CRAN (R 3.4.4)
broom 0.5.0 2018-07-17 CRAN (R 3.4.1)
BSgenome * 1.44.2 2017-09-23 Bioconductor
cellranger 1.1.0 2016-07-27 CRAN (R 3.4.0)
cli 1.0.0 2017-11-05 CRAN (R 3.4.2)
clue 0.3-55 2018-04-23 CRAN (R 3.4.4)
cluster * 2.0.6 2017-03-10 CRAN (R 3.4.1)
coda 0.19-1 2016-12-08 cran (@0.19-1)
codetools 0.2-15 2016-10-05 CRAN (R 3.4.1)
coin 1.2-2 2017-11-28 CRAN (R 3.4.3)
colorspace 1.3-2 2016-12-14 CRAN (R 3.4.0)
compiler 3.4.1 2017-07-07 local
cowplot * 0.9.3 2018-07-15 CRAN (R 3.4.4)
crayon 1.3.4 2017-09-16 CRAN (R 3.4.1)
data.table 1.11.4 2018-05-27 CRAN (R 3.4.4)
datasets * 3.4.1 2017-07-07 local
DBI 1.0.0 2018-05-02 CRAN (R 3.4.4)
DelayedArray 0.2.7 2017-06-03 Bioconductor
devtools * 1.13.6 2018-06-27 CRAN (R 3.4.4)
digest 0.6.16 2018-08-22 CRAN (R 3.4.4)
dndscv * 0.0.0.9 2018-07-23 Github (im3sanger/dndscv@88d5d69)
doParallel 1.0.11 2017-09-28 CRAN (R 3.4.2)
dplyr * 0.7.6 2018-06-29 CRAN (R 3.4.4)
emmeans 1.3.0 2018-10-26 CRAN (R 3.4.4)
estimability 1.3 2018-02-11 CRAN (R 3.4.3)
evaluate 0.11 2018-07-17 CRAN (R 3.4.4)
fansi 0.3.0 2018-08-13 CRAN (R 3.4.4)
forcats * 0.3.0 2018-02-19 CRAN (R 3.4.3)
foreach 1.4.4 2017-12-12 CRAN (R 3.4.3)
foreign 0.8-69 2017-06-22 CRAN (R 3.4.1)
GenomeInfoDb * 1.12.3 2017-10-05 Bioconductor
GenomeInfoDbData 0.99.0 2017-08-24 Bioconductor
GenomicAlignments 1.12.2 2017-08-19 Bioconductor
GenomicFeatures 1.28.5 2017-09-20 Bioconductor
GenomicRanges * 1.28.6 2017-10-04 Bioconductor
ggeffects 0.7.0 2018-11-17 CRAN (R 3.4.4)
ggplot2 * 3.0.0 2018-07-03 CRAN (R 3.4.4)
ggridges 0.5.0 2018-04-05 CRAN (R 3.4.4)
glmmTMB 0.2.2.0 2018-07-03 CRAN (R 3.4.4)
glue 1.3.0 2018-07-17 CRAN (R 3.4.4)
graphics * 3.4.1 2017-07-07 local
grDevices * 3.4.1 2017-07-07 local
grid 3.4.1 2017-07-07 local
gridBase 0.4-7 2014-02-24 CRAN (R 3.4.0)
gridExtra * 2.3 2017-09-09 CRAN (R 3.4.1)
gtable 0.2.0 2016-02-26 CRAN (R 3.4.0)
haven 1.1.2 2018-06-27 CRAN (R 3.4.4)
hdp * 0.1.5 2018-05-01 Github (nicolaroberts/hdp@c78989b)
hms 0.4.2 2018-03-10 CRAN (R 3.4.4)
htmltools 0.3.6 2017-04-28 CRAN (R 3.4.0)
httr 1.3.1 2017-08-20 CRAN (R 3.4.1)
IRanges * 2.10.5 2017-10-08 Bioconductor
iterators 1.0.10 2018-07-13 CRAN (R 3.4.4)
jsonlite 1.5 2017-06-01 CRAN (R 3.4.0)
kableExtra * 0.9.0 2018-05-21 CRAN (R 3.4.4)
knitr * 1.20 2018-02-20 CRAN (R 3.4.3)
labeling 0.3 2014-08-23 CRAN (R 3.4.0)
lattice 0.20-35 2017-03-25 CRAN (R 3.4.1)
lazyeval 0.2.1 2017-10-29 CRAN (R 3.4.2)
lme4 * 1.1-18-1 2018-08-17 CRAN (R 3.4.4)
lmerTest * 3.0-1 2018-04-23 CRAN (R 3.4.4)
lubridate 1.7.4 2018-04-11 CRAN (R 3.4.4)
magrittr * 1.5 2014-11-22 CRAN (R 3.4.0)
MASS * 7.3-47 2017-02-26 CRAN (R 3.4.1)
Matrix * 1.2-10 2017-05-03 CRAN (R 3.4.1)
matrixStats 0.54.0 2018-07-23 CRAN (R 3.4.4)
memoise 1.1.0 2017-04-21 CRAN (R 3.4.0)

methods * 3.4.1 2017-07-07 local
minqa 1.2.4 2014-10-09 CRAN (R 3.4.0)
mnormt 1.5-5 2016-10-15 CRAN (R 3.4.0)
modelr 0.1.2 2018-05-11 CRAN (R 3.4.4)
modeltools 0.2-22 2018-07-16 CRAN (R 3.4.4)
multcomp 1.4-8 2017-11-08 CRAN (R 3.4.2)
3
nature research | reporting summary October 2018
munsell 0.5.0 2018-06-12 CRAN (R 3.4.4)
MutationalPatterns * 1.2.1 2017-05-09 Bioconductor
mvtnorm 1.0-8 2018-05-31 CRAN (R 3.4.4)
nlme 3.1-131 2017-02-06 CRAN (R 3.4.1)
nloptr 1.0.4 2014-08-04 CRAN (R 3.4.0)
NMF * 0.21.0 2018-03-06 CRAN (R 3.4.4)
numDeriv 2016.8-1 2016-08-27 CRAN (R 3.4.0)
parallel * 3.4.1 2017-07-07 local
pheatmap * 1.0.10 2018-05-19 CRAN (R 3.4.4)
pillar 1.3.0 2018-07-14 CRAN (R 3.4.4)
pkgconfig 2.0.2 2018-08-16 CRAN (R 3.4.4)
pkgmaker * 0.27 2018-05-25 CRAN (R 3.4.4)
plyr 1.8.4 2016-06-08 CRAN (R 3.4.0)
pracma 2.1.5 2018-08-25 CRAN (R 3.4.4)
prediction 0.3.6.1 2018-12-04 CRAN (R 3.4.1)
psych 1.8.4 2018-05-06 CRAN (R 3.4.4)
purrr * 0.2.5 2018-05-29 CRAN (R 3.4.4)
pwr 1.2-2 2018-03-03 CRAN (R 3.4.3)
R6 2.2.2 2017-06-17 CRAN (R 3.4.0)
RColorBrewer * 1.1-2 2014-12-07 CRAN (R 3.4.0)
Rcpp 0.12.18 2018-07-23 CRAN (R 3.4.4)
RCurl 1.95-4.11 2018-07-15 CRAN (R 3.4.4)
readr * 1.1.1 2017-05-16 CRAN (R 3.4.0)
readxl * 1.1.0 2018-04-20 CRAN (R 3.4.4)
registry * 0.5 2017-12-03 CRAN (R 3.4.3)
reshape2 1.4.3 2017-12-11 CRAN (R 3.4.3)
rlang * 0.2.2 2018-08-16 CRAN (R 3.4.4)
rmarkdown 1.10 2018-06-11 CRAN (R 3.4.4)
rngtools * 1.3.1 2018-05-15 CRAN (R 3.4.4)
rprojroot 1.3-2 2018-01-03 CRAN (R 3.4.1)
Rsamtools 1.28.0 2017-04-25 Bioconductor
RSQLite 2.1.1 2018-05-06 CRAN (R 3.4.4)
rstudioapi 0.7 2017-09-07 CRAN (R 3.4.1)
rtracklayer * 1.36.6 2017-10-12 Bioconductor
rvest 0.3.2 2016-06-17 CRAN (R 3.4.0)
S4Vectors * 0.14.7 2017-10-08 Bioconductor
sandwich 2.5-0 2018-08-17 CRAN (R 3.4.4)
scales 1.0.0 2018-08-09 CRAN (R 3.4.4)
seqinr * 3.4-5 2017-08-01 CRAN (R 3.4.1)
sjlabelled 1.0.15 2018-11-22 CRAN (R 3.4.4)
sjmisc 2.7.6 2018-11-06 CRAN (R 3.4.4)
sjPlot 2.6.1 2018-10-14 CRAN (R 3.4.4)
sjstats 0.17.2 2018-11-15 CRAN (R 3.4.4)
snakecase 0.9.2 2018-08-14 CRAN (R 3.4.4)
splines 3.4.1 2017-07-07 local
stats * 3.4.1 2017-07-07 local
stats4 * 3.4.1 2017-07-07 local
stringdist 0.9.5.1 2018-06-08 CRAN (R 3.4.4)
stringi 1.2.4 2018-07-20 CRAN (R 3.4.4)
stringr * 1.3.1 2018-05-10 CRAN (R 3.4.4)
SummarizedExperiment 1.6.5 2017-09-29 Bioconductor
survival 2.41-3 2017-04-04 CRAN (R 3.4.1)
TH.data 1.0-9 2018-07-10 CRAN (R 3.4.4)
tibble * 1.4.2 2018-01-22 CRAN (R 3.4.3)
tidyr * 0.8.1 2018-05-18 CRAN (R 3.4.4)
tidyselect 0.2.4 2018-02-26 CRAN (R 3.4.3)
tidyverse * 1.2.1 2017-11-14 CRAN (R 3.4.2)
TMB 1.7.15 2018-11-09 CRAN (R 3.4.4)
tools 3.4.1 2017-07-07 local
utf8 1.1.4 2018-05-24 CRAN (R 3.4.4)
utils * 3.4.1 2017-07-07 local
VariantAnnotation 1.22.3 2017-06-24 Bioconductor
viridisLite 0.3.0 2018-02-01 CRAN (R 3.4.3)
withr 2.1.2 2018-03-15 CRAN (R 3.4.4)
XML 3.98-1.16 2018-08-19 CRAN (R 3.4.4)
xml2 1.2.0 2018-01-24 CRAN (R 3.4.3)
xtable 1.8-3 2018-08-29 CRAN (R 3.4.1)

XVector * 0.16.0 2017-04-25 Bioconductor
zlibbioc 1.22.0 2017-04-25 Bioconductor
zoo 1.8-3 2018-07-16 CRAN (R 3.4.4)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All of the whole genome sequencing data analysed in this study have been submitted to the European Genome-Phenome Archive with accession number 716 EGAS00001002471

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal prior sample size calculation was performed. Sample size was chosen to give good representation of inter-patient variability in mutation burden.
Data exclusions	Samples with sequencing depth <15x were excluded from the final analysis. The coverage cut-off was selected based on the probability of calling a variant with a high sensitivity.
Replication	A set of biological 'near-replicates' was obtained for 18 glands. For these experiments, we collected two samples from the same endometrial gland which was identified on two or more consecutive levels using z-stacking approach; each sample was processed separately with an independent DNA extraction, library preparation and whole genome sequencing.
Randomization	This sequencing study included samples of normal endometrial tissue only; there was no randomization or comparison to a control group.
Blinding	This study included samples of normal endometrial tissue only; no blinding was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	This study includes samples of normal endometrium from 28 women, aged 19 to 81.

Recruitment	Human endometrium is a dynamic tissue that adopts various physiological states during life including pre-menarche, menstrual cycling, pregnancy and post-menopause. We therefore sought to access samples from as wide an age range of women as possible through access to different cohorts. 1. Transplant organ donors and post-mortem samples (women aged 19-81 years) The transplant organ donors were individuals who had suffered serious accidents that would inevitably lead to death and therefore, while still alive, organs such as kidney, liver, lung and heart were removed for live transplants into individuals in whom one of these organs was failing, with additional samples being taken for research purposes (including the studies reported here). The transplant organ donor and post-mortem samples of endometrium were taken from individuals aged 19 to 81 years who therefore died of non-gynaecological causes, without known gynaecological symptoms or clinical history, and without histological abnormality of gynaecological organs on microscopical examination. This cohort represents the most optimal approach to sampling and investigating normal endometrium, as gynaecological symptomatology played no part at all in the ascertainment of the individuals and there was no clinical or pathological evidence of abnormality. The samples from the transplant organ donors were collected with an informed consent obtained from donor's family (REC reference: 15/EE/0152 NRES Committee East of England – Cambridge South). The post-mortem samples were collected at autopsy following death from non-gynaecological causes; the use of this material was approved by the London, Surrey Research Ethics Committee (REC reference 17/LO/1801, 26/10/2017) and East of Scotland Research Ethics Service (REC reference: 17/ES/0102, 27/07/2017). 2. Infertility donors (women in their 20's-40's) To ensure that the study incorporates a sufficient number of women in their 20's-40's, we also sought to include biopsies from individuals attending an infertility centre. This source of samples provides a unique opportunity to obtain endometrial samples from relatively young women (other than the rare cases of autopsy and transplant donors described above), something that has also been recognised by other groups working in this field with similar sets of samples used in recent publications on the genomics of normal endometrium. In the particular centre from which we obtained samples women are seen for a range of reasons, including disorders of the fallopian tubes (tubal obstruction/pelvic inflammatory disease secondary to Chlamydia infection), miscarriage, endometriosis, congenital anatomical uterine abnormalities, polycystic ovary syndrome (PCOS) as well as conditions related to the male partner, including sperm disorders. All the studied endometrial biopsies were examined by two histopathologists who confirmed that they were histologically normal. Informed consent was obtained and biopsies collected and stored at the Arden Tissue Bank, University Hospitals Coventry and Warwickshire NHS Trust in line with the protocols approved by the NRES Committee South Central Southampton B (REC reference 12/SC/0526, 19/04/2013). The main potential bias with this cohort is the fact that we would only obtain samples from women of reproductive age, which is why we sought to find additional sources. 3. Hysterectomy samples (women aged 47 and 49 years) We also included samples of normal endometrium from women who underwent total abdominal hysterectomy for benign non-endometrial pathologies. Biopsies were collected, snap frozen and stored at the Human Research Tissue Bank, Cambridge University Hospitals NHS Foundation Trust in line with the protocols approved by the NRES Committee East of England (REC reference 11/EE/0011, 11/03/2011). One of the main potential biases with this collection is that while hysterectomy is a relatively standard surgical procedure, it is usually avoided in women under the age 40 and is often not the treatment of choice in those who have not yet completed their families. Therefore, a study focused on such samples alone would've hindered us from gaining insights into the somatic evolution occurring in the endometrium of younger women.
Ethics oversight	As outlined above

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

AIM2 inflammasome surveillance of DNA damage shapes neurodevelopment

https://doi.org/10.1038/s41586-020-2174-3

Received: 24 May 2018

Accepted: 10 February 2020

Published online: 8 April 2020

Check for updates

Catherine R. Lammert[1,2], Elizabeth L. Frost[1], Calli E. Bellinger[1], Ashley C. Bolte[1,3,4], Celia A. McKee[1], Mariah E. Hurt[1], Matt J. Paysour[1], Hannah E. Ennerfelt[1,2] & John R. Lukens[1,2]✉

Neurodevelopment is characterized by rapid rates of neural cell proliferation and differentiation followed by massive cell death in which more than half of all recently generated brain cells are pruned back. Large amounts of DNA damage, cellular debris, and by-products of cellular stress are generated during these neurodevelopmental events, all of which can potentially activate immune signalling. How the immune response to this collateral damage influences brain maturation and function remains unknown. Here we show that the AIM2 inflammasome contributes to normal brain development and that disruption of this immune sensor of genotoxic stress leads to behavioural abnormalities. During infection, activation of the AIM2 inflammasome in response to double-stranded DNA damage triggers the production of cytokines as well as a gasdermin-D-mediated form of cell death known as pyroptosis[1–4]. We observe pronounced AIM2 inflammasome activation in neurodevelopment and find that defects in this sensor of DNA damage result in anxiety-related behaviours in mice. Furthermore, we show that the AIM2 inflammasome contributes to central nervous system (CNS) homeostasis specifically through its regulation of gasdermin-D, and not via its involvement in the production of the cytokines IL-1 and/or IL-18. Consistent with a role for this sensor of genomic stress in the purging of genetically compromised CNS cells, we find that defective AIM2 inflammasome signalling results in decreased neural cell death both in response to DNA damage-inducing agents and during neurodevelopment. Moreover, mutations in AIM2 lead to excessive accumulation of DNA damage in neurons as well as an increase in the number of neurons that incorporate into the adult brain. Our findings identify the inflammasome as a crucial player in establishing a properly formed CNS through its role in the removal of genetically compromised cells.

Here we asked whether the damage signals generated during neurodevelopment trigger activation of the innate immune response and, if so, how this immune activation shapes neurodevelopment and behaviour. The high levels of replicative stress and cell death that occur during brain maturation are known to generate several damage or danger signals such as DNA damage, ATP and mitochondrial stress, all of which can trigger inflammasome activation. Inflammasomes are multiprotein complexes that generally consist of an intracellular receptor such as NLRP3 or AIM2, the adaptor protein ASC, and the enzyme caspase-1. Formation of this innate immune signalling platform coordinates the activation of caspase-1, which can subsequently promote the production of IL-1 and IL-18, as well as a gasdermin-D-mediated form of cell death commonly referred to as pyroptosis.

ASC specks form in neurodevelopment

Because both cytokine production and cell death have been shown to be pivotal modulators of neurodevelopment[5–14], along with the fact that key components of the inflammasome are highly expressed during developmental time points (Extended Data Fig. 1a–c), we sought to determine whether inflammasome activation influences brain maturation and CNS function. We first asked whether inflammasome activation is observed during neurodevelopment. To test this, we used ASC reporter mice to track the development of ASC specks, which are prototypical markers of inflammasome activation[15]. Notably, in the developing brain at postnatal day 5 (P5)—a time point characterized by high levels of DNA damage and cell death[16,17]—we observed high levels of ASC speck formation throughout the brain (Fig. 1a, b, Extended Data Fig. 2a). By contrast, we detected very few ASC specks in fully matured lymphoid organs such as the lymph nodes (Fig. 1b, Extended Data Fig. 2b, c).

Inflammasomes influence behaviour

To assess the importance of inflammasome activation in establishing a properly functioning CNS, we performed a range of behavioural tests

[1]Center for Brain Immunology and Glia (BIG), Department of Neuroscience, School of Medicine, University of Virginia, Charlottesville, VA, USA. [2]Neuroscience Graduate Program, School of Medicine, University of Virginia, Charlottesville, VA, USA. [3]Medical Scientist Training Program, School of Medicine, University of Virginia, Charlottesville, VA, USA. [4]Immunology Training Program, School of Medicine, University of Virginia, Charlottesville, VA, USA. ✉e-mail: jrl7n@virginia.edu

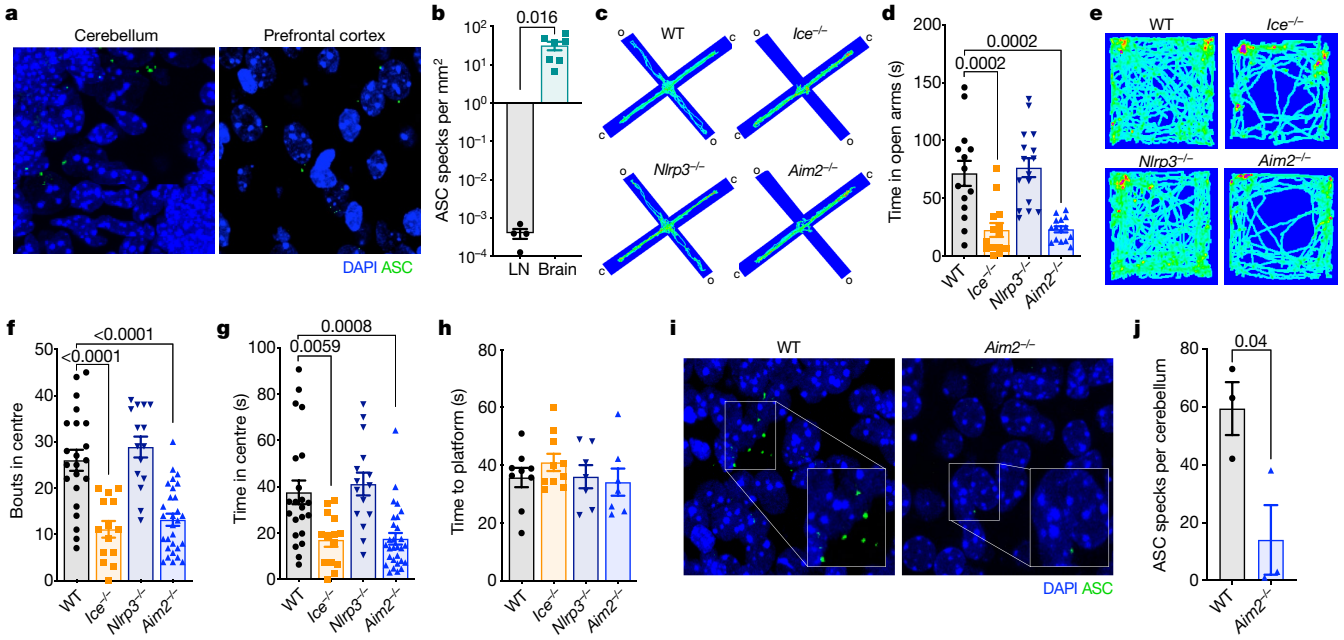


Fig. 1 | Inflammasome activation occurs in the CNS during neurodevelopment and disruption of the AIM2 inflammasome results in anxiety-related behaviours. a, Brains from P5 ASC-citrine reporter mice were analysed for the presence of ASC specks (green). Images are representative of three independent experiments with similar results. Original magnification, ×40. **b**, Number of ASC specks formed in mice brains at P5 ($n = 7$ mice; from three independent experiments) and in deep cervical lymph nodes (LN) of adult mice (8–16 weeks old) ($n = 4$; from two independent experiments). **c**–**h**, Wild-type (WT), *Ice*$^{-/-}$, *Nlrp3*$^{-/-}$ and *Aim2*$^{-/-}$ mice (8–12 weeks old) were assessed for behavioural abnormalities. Anxiety behaviours were assessed using the elevated plus maze and the open-field test (**c**–**g**). **c**, Representative heat maps showing the average time spent in the open (o) and closed (c) arms of the elevated plus maze. Heat maps are from six independent experiments with similar results. **d**, Quantification of the time spent in the open arms of the elevated plus maze (WT $n = 14$, *Ice*$^{-/-}$ $n = 14$, *Nlrp3*$^{-/-}$ $n = 16$, *Aim2*$^{-/-}$ $n = 15$; from three independent experiments). **e**, Representative heat maps of the activity of mice in the open-field test. Heat maps are from six independent experiments with similar results. **f**, **g**, Quantification of the bouts into (**f**) and the time spent in (**g**) the centre of the open-field arena (WT $n = 22$, *Ice*$^{-/-}$ $n = 14$, *Nlrp3*$^{-/-}$ $n = 15$, *Aim2*$^{-/-}$ $n = 29$; from five independent experiments). **h**, Visual platform test of the Morris water maze (WT $n = 9$, *Ice*$^{-/-}$ $n = 10$, *Nlrp3*$^{-/-}$ $n = 7$, *Aim2*$^{-/-}$ $n = 7$; from two independent experiments). **i**, **j**, Cerebellar ASC speck formation in wild-type and *Aim2*$^{-/-}$ P5 mice. **i**, Images are representative of two independent experiments with similar results. **j**, Quantification of cerebellar ASC speck formation in wild-type ($n = 3$) and *Aim2*$^{-/-}$ ($n = 3$) mice. Original magnification, ×40. Data are representative of two independent experiments. All n values refer to the number of mice used. Data are mean ± s.e.m. P values were calculated by unpaired two-tailed Student's t-test (**b**, **j**) or one-way analysis of variance (ANOVA) with Tukey's post hoc tests (**d**, **f**–**h**).

on mice deficient in both *Casp1* and *Casp11*, which are also referred to as *Ice*$^{-/-}$ mice (note *Casp11* is also known as *Casp4*). Genetic ablation of the inflammasome resulted in profound anxiety-like behaviours in the elevated plus maze and in open-field testing (Fig. 1c–g). More specifically, we observed that *Ice*$^{-/-}$ mice spent significantly less time exploring the open arm of the elevated plus maze (Fig. 1c, d). Moreover, in open-field testing, *Ice*$^{-/-}$ mice explored the centre significantly less than wild-type mice, urinated more frequently, and produced more faecal pellets (Fig. 1e–g, Extended Data Fig. 3a, b). To ensure that impaired vision and/or locomotor activity did not underlie the poor performance of the inflammasome-deficient mice in our behavioural tests, we evaluated their ability to find a visible escape platform in the Morris water maze. In these studies, deletion of the inflammasome did not negatively affect the ability of mice to reach the visual platform, indicating that neither impaired vision nor motor deficits are likely to contribute to the differences in performance seen in our elevated plus maze and open-field tests (Fig. 1h). Genetic ablation of inflammasome signalling in mice deficient in caspase-1 and caspase-11 also did not result in global behavioural abnormalities, as *Ice*$^{-/-}$ mice performed normally in the tail suspension and sucrose preference tests (Extended Data Fig. 3c, d), both of which are commonly used to assess depressive behaviours. Collectively, these results indicate that impaired activation of the inflammasome leads to behavioural abnormalities that include the development of pronounced anxiety-like behaviours.

The immune system is equipped with a repertoire of intracellular receptors that enable the host to coordinate inflammasome activation in response to a diverse array of pathogens and endogenous damage or danger signals. We first turned our attention to a potential role for NLRP3 in our model, as it is known to induce inflammasome activation in response to a diverse array of damage/danger-associated molecular patterns (DAMPs) that are probably generated during normal brain maturation (such as ATP, damaged mitochondria, and reactive oxygen species)[4]. Notably, *Nlrp3*$^{-/-}$ mice performed similarly to wild-type mice in the elevated plus maze, open-field test, visual platform test, and depressive assays (Fig. 1c–h, Extended Data Fig. 3). These results suggest that NLRP3 does not coordinate the inflammasome activation needed to prevent the development of anxiety-like behaviours in mice.

The AIM2 inflammasome affects behaviour

Maintenance of genomic integrity is essential for CNS health and mounting evidence suggests that the inability to control genotoxic stress centrally contributes to several neurodevelopmental, psychiatric and neurodegenerative disorders[18,19]. In most cases, DNA damage is quickly remediated by repair pathways. However, DNA insults can persist as a result of unsuccessful repair attempts and/or impaired removal of DNA damage[20,21]. The sensing of DNA damage by AIM2 in peripheral immune cells can trigger inflammasome activation[22]. In addition, previous work has described roles for AIM2 in models of CNS injury[23,24] and has also begun to characterize how deletion of AIM2 can alter neuronal morphology and influence behaviour[25]. However, whether the AIM2 inflammasome is activated during neurodevelopment and, if so, how

this affects brain maturation and behaviour remain unclear. When we tested AIM2-deficient mice for anxiety-like behaviours we found that *Aim2*$^{-/-}$ mice phenocopied *Ice*$^{-/-}$ mice and displayed anxiety-like behaviours in both the elevated plus maze and the open-field test (Fig. 1c–g, Extended Data Fig. 3a, b). Like *Ice*$^{-/-}$ mice, AIM2-deficient mice reached the visual platform in the Morris water maze in similar times to wild-type controls (Fig. 1h) and performed normally in both the tail suspension and sucrose preference tests (Extended Data Fig. 3c, d).

DNA damage and AIM2 promote ASC specks

Because key AIM2 inflammasome-associated genes are abundantly expressed during neurodevelopment (Extended Data Fig. 1) and double-stranded DNA (dsDNA) is known to activate AIM2, we sought to determine whether the DNA damage that normally arises during neurodevelopment can trigger activation of the AIM2 inflammasome in the developing brain. First, we explored whether the inflammasome activation observed during neurodevelopment occurs in close proximity to cells containing DNA damage. To this end, we evaluated the spatial localization of ASC specks in relation to cells that co-stain for the markers of DNA damage γH2AX and 53BP1. Many of the ASC specks formed in the developing brain were found to be in the vicinity of cells that co-stained for γH2AX and 53BP1 (Extended Data Fig. 4a). Moreover, the bulk of this inflammasome activation at P5 in neurodevelopment is dependent on AIM2 surveillance, as genetic abrogation of AIM2 greatly decreased the number of ASC specks detected in the developing cerebellum (Fig. 1i, j). To further interrogate whether the AIM2 inflammasome can be activated in the developing brain in response to DNA damage, we induced overt DNA damage at P5 in the brains of wild-type and AIM2-deficient mice by exposing them to ionizing radiation. Exposure to ionizing radiation resulted in increased expression of DNA damage markers in wild-type mice (that is, γH2AX staining) and this corresponded to concomitant increases in inflammasome activation (ASC speck formation) (Extended Data Fig. 4b–e). We observed a corresponding increase in γH2AX staining in AIM2-deficient mice treated with ionizing radiation (Extended Data Fig. 4b, c). However, the formation of ASC specks was substantially blunted in the absence of AIM2 (Extended Data Fig. 4d, e). Together, these findings suggest that activation of the AIM2 inflammasome is likely to occur during neurodevelopment in response to DNA damage, and that disruptions in this pathway can lead to behavioural abnormalities.

Gasdermin-D shapes behaviour

Inflammasome activation can lead to the secretion of IL-1 and IL-18, and to a gasdermin-D-mediated form of cell death—both of which can potentially affect neurodevelopment and behaviour. Cytokines have been shown to be pivotal modulators of neurodevelopment, CNS function, and behaviour[5,6,8–10]. In particular, the inflammasome-derived cytokines IL-1 and IL-18 have been reported to have especially prominent effects on the CNS[6,8–10]. Therefore, we sought to determine whether the observed behavioural abnormalities in AIM2 inflammasome-deficient mice were caused by disruptions in AIM2 inflammasome-mediated production of IL-1 and/or IL-18. Notably, the abrogation of IL-1 or IL-18 signalling did not promote anxiety-related phenotypes (Fig. 2a, b, Extended Data Fig. 5a–c). To interrogate whether IL-1 and IL-18 can have compensatory roles in shaping behaviour, we also evaluated anxiety-related phenotypes in mice that lack MYD88, which is an essential adaptor molecule required for both IL-1R and IL-18R signalling. However, abrogation of MYD88 signalling did not influence performance in tests measuring anxiety-related behaviours (Fig. 2a, b, Extended Data Fig. 5a–e).

In addition to orchestrating the production of IL-1 and IL-18, activation of the AIM2 inflammasome can also incite gasdermin-D-mediated cell death. To explore the role that gasdermin-D has in driving anxiety-like phenotypes, we assessed the performance of gasdermin-D

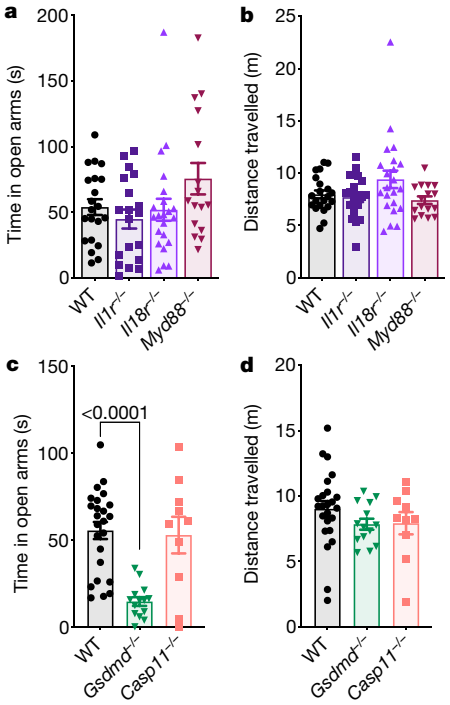


Fig. 2 | Lack of gasdermin-D activation drives anxiety-like behaviours in mice. a–d, Anxiety-related behaviours were assessed in adult (8–12 weeks old) wild-type, *Il1r*$^{-/-}$ (also known as *Il1r1*$^{-/-}$), *Il18r*$^{-/-}$ (*Il18r1*$^{-/-}$), *Myd88*$^{-/-}$, *Gsdmd*$^{-/-}$ and *Casp11*$^{-/-}$ mice using the elevated plus maze. **a**, **b**, Quantification of the time spent in the open arms (**a**) and the distance travelled (**b**) (WT $n = 21$, *Il1r*$^{-/-}$ $n = 19$, *Il18r*$^{-/-}$ $n = 22$, *Myd88*$^{-/-}$ $n = 16$; from four independent experiments) during the elevated plus maze test. **c**, **d**, Time spent in the open arms (**c**) and the total distance travelled (**d**) (WT $n = 24$, *Gsdmd*$^{-/-}$ $n = 14$, *Casp11*$^{-/-}$ $n = 10$; from two independent experiments) during the elevated plus maze test. All *n* values refer to the number of mice used. Data are mean ± s.e.m. Statistics were calculated by one-way ANOVA with Tukey's post hoc tests.

knockout mice in the elevated plus maze. Similar to *Ice*$^{-/-}$ and *Aim2*$^{-/-}$ mice (Fig. 1c, d), *Gsdmd*$^{-/-}$ mice spent less time exploring the open arm of the elevated plus maze (Fig. 2c, d). Caspase-11, like caspase-1, can also orchestrate gasdermin-D activation through noncanonical inflammasome signalling[26]. Nevertheless, caspase-11-deficient mice performed similarly to wild-type mice in both the elevated plus maze and open-field testing (Fig. 2c, d, Extended Data Fig. 5f–h). Together, these results suggest that the behavioural abnormalities observed in AIM2 inflammasome-deficient mice are probably not due to defects in caspase-1-mediated production of IL-1 and/or IL-18, but instead result from impaired gasdermin-D signalling.

AIM2 coordinates neural cell death

More than half of all neural cells are eliminated during neurodevelopment[11]. This process of neural cell pruning has beneficial roles in the sculpting of strong connections in the brain and, consistent with this idea, disruptions in CNS cell death during development have been shown to cause neurological dysfunction[11–14]. Neuronal dieback is thought to occur solely through apoptotic cell death; however, this requires revisiting with the discovery of other forms of programmed cell death that include pyroptosis, necroptosis and autophagic cell death[27]. Given our data indicating that the cell death executioner gasdermin-D and the DNA damage sensor AIM2 both have key roles in limiting neurological dysfunction, we speculated that AIM2 inflammasome-induced cell death may help to prevent genetically compromised cells from being incorporated into the mature brain. To explore this

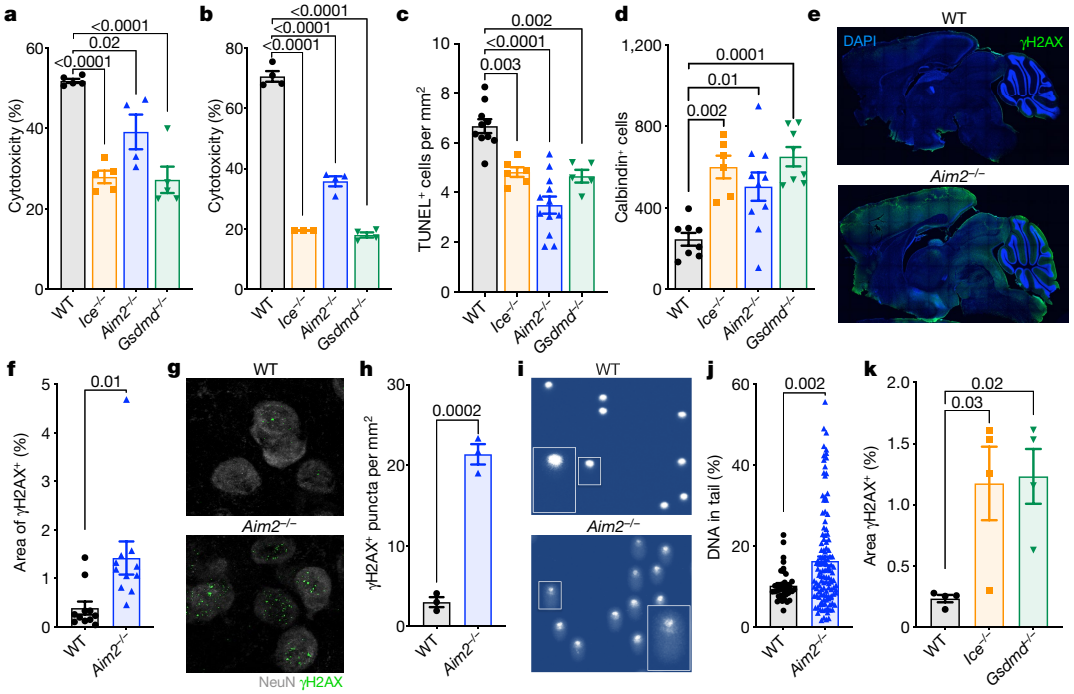


Fig. 3 | Activation of the AIM2 inflammasome in response to DNA damage coordinates CNS cell death and limits the accumulation of DNA damage in the brain. a, **b**, Mixed neural cultures from mice at P0 were primed with lipopolysaccharide (LPS) for 4 h followed by overnight treatment with either 40 Gy of ionizing radiation (WT $n = 5$, *Ice*$^{-/-}$ $n = 5$, *Aim2*$^{-/-}$ $n = 4$, *Gsdmd*$^{-/-}$ $n = 5$) (**a**) or 100 μM etoposide (WT $n = 4$, *Ice*$^{-/-}$ $n = 3$, *Aim2*$^{-/-}$ $n = 4$, *Gsdmd*$^{-/-}$ $n = 4$) (**b**). Cell death was measured by the release of lactate dehydrogenase (LDH). Data are representative of three independent experiments. **c**, Quantification of cerebellar TUNEL staining in P5 mice (WT $n = 10$, *Ice*$^{-/-}$ $n = 6$, *Aim2*$^{-/-}$ $n = 12$, *Gsdmd*$^{-/-}$ $n = 5$; from three independent experiments). **d**, Enumeration of cerebellar calbindin^{+} Purkinje neurons in adult (8–12 weeks old) mice (WT $n = 8$, *Ice*$^{-/-}$ $n = 6$, *Aim2*$^{-/-}$ $n = 10$, *Gsdmd*$^{-/-}$ $n = 8$; from three independent experiments). **e**, **f**, Adult brains were evaluated for levels of DNA damage (γH2AX, green). **e**, Images are representative of three independent experiments with similar results. Original magnification, ×10. **f**, Quantification of γH2AX staining in adult mice (WT $n = 11$, *Aim2*$^{-/-}$ $n = 11$; from three independent experiments). **g**, **h**, Adult brains were evaluated for γH2AX staining in NeuN-expressing neurons in the amygdala. **g**, Images are representative of two independent experiments with similar results. Original magnification, ×63. **h**, Enumeration of γH2AX puncta in the amygdala (WT $n = 3$, *Aim2*$^{-/-}$ $n = 3$; data are representative of two independent experiments). **i**, **j**, DNA damage was evaluated in the cortex of 10-week-old wild-type and *Aim2*$^{-/-}$ mice by comet assay. **i**, Representative images of single-cell electrophoresis gels from three independent experiments with similar results. Original magnification, ×40. **j**, Quantification of the percentage of DNA in the tail (WT $n = 38$, *Aim2*$^{-/-}$ $n = 120$; from three independent experiments). **k**, Quantification of γH2AX staining in sagittal brain sections from wild-type ($n = 4$), *Ice*$^{-/-}$ ($n = 4$) and *Gsdmd*$^{-/-}$ ($n = 4$) mice. Data are from two independent experiments. *n* values refer to biological replicates from representative experiments (**a**, **b**, **j**) or to the number of mice used (**c**, **d**, **f**, **h**, **k**). Data are mean ± s.e.m. Statistics were calculated by one-way ANOVA with Tukey's post hoc tests (**a**–**d**, **k**) and unpaired two-tailed Student's *t*-test (**f**, **h**, **j**).

possibility, we first evaluated whether AIM2 inflammasome signalling is involved in CNS cell turnover in response to endogenous DNA damage. Mixed cortical neural cells from wild-type, *Aim2*$^{-/-}$, *Ice*$^{-/-}$ and *Gsdmd*$^{-/-}$ mice were either transfected with dsDNA (poly(dA:dT)) as a positive control or treated with ionizing radiation or the topoisomerase II inhibitor etoposide to induce endogenous DNA damage. We detected a substantial reduction in DNA damage-induced cell death in CNS cells lacking AIM2, caspase-1 and caspase-11, or gasdermin-D (Fig. 3a, b, Extended Data Fig. 6a, b).

To investigate whether the AIM2 inflammasome is involved in coordinating neural cell dieback in vivo, we evaluated cell death in the cerebellum of wild-type and AIM2 inflammasome-deficient mice at P5, as this brain region has been shown to undergo DNA damage-induced cell death at this time point in neurodevelopment[16,17]. Genetic ablation of AIM2 resulted in reduced levels of cell death, as indicated by a decrease in TUNEL and propidium iodide staining in the cerebellums of *Aim2*$^{-/-}$ P5 mice (Fig. 3c, Extended Data Fig. 7a–c). A reduction in cell death was also seen in cerebellums from *Ice*$^{-/-}$ and *Gsdmd*$^{-/-}$ P5 mice, indicating that the AIM2 inflammasome and gasdermin-D are involved in orchestrating cell death at P5 in the developing brain (Fig. 3c, Extended Data Fig. 7a). Notably, disruptions in AIM2 inflammasome signalling did not completely abrogate levels of CNS cell death at this time point, which suggests that other forms of cell death such as apoptosis are also contributing to the pruning of CNS cells in AIM2 inflammasome-deficient mice. To further test our working model, which proposes that AIM2 inflammasome signalling coordinates the removal of CNS cells containing DNA damage, we induced overt DNA damage in the developing brains of wild-type and AIM2-deficient mice by exposing them to ionizing radiation.Exposure to ionizing radiation resulted in increased TUNEL staining in wild-type and *Aim2*$^{-/-}$ mice; however, cell death induced by ionizing radiation was substantially blunted in the absence of AIM2 (Extended Data Fig. 7d, e). These findings suggest that AIM2 can execute cell death in response to ionizing-radiation-driven DNA damage.

AIM2 limits DNA damage levels in the CNS

If the AIM2 inflammasome does function in the purging of genetically compromised cells from the brain, we would expect that disruptions in this pathway would cause greater incorporation of cells into the adult brain as well as increased levels of DNA damage. Consistent with this idea, we observed increased numbers of calbindin^{+} Purkinje neurons in the brains of mice lacking AIM2, caspase-1 and caspase-11, or gasdermin-D (Fig. 3d, Extended Data Fig. 8). Furthermore, we also detected

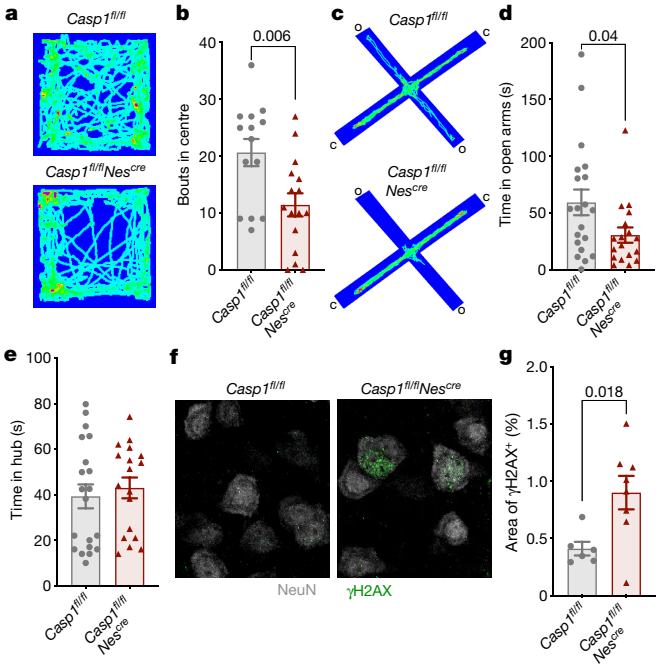


Fig. 4 | CNS-specific deletion of caspase-1 results in anxiety-like behaviours and DNA damage accumulation in the brain. a–**e**, Anxiety-associated behaviours were assessed in adult (8–12 weeks old) *Casp1*$^{fl/fl}$ and *Casp1*$^{fl/fl}$*Nes*^{cre} mice. **a**, Representative heat maps of the path mice travelled in the open-field arena; from three independent experiments with similar results. **b**, Quantification of bouts into the centre of the open-field arena (*Casp1*$^{fl/fl}$ $n = 14$, *Casp1*$^{fl/fl}$*Nes*^{cre} $n = 16$; from three independent experiments). **c**, Representative heat maps depicting path of travel through open and closed arms of the elevated plus maze. Data are from four independent experiments with similar results. **d**, Quantification of time spent in the open arms of the elevated plus maze (**d**) and time in the hub (**e**) (*Casp1*$^{fl/fl}$ $n = 20$, *Casp1*$^{fl/fl}$*Nes*^{cre} $n = 18$). Data are from four independent experiments. **f**, **g**, Adult brains were evaluated for levels of DNA damage (γH2AX, green) in NeuN-expressing neurons. **f**, Representative cortex images from two independent experiments with similar results. Original magnification, ×63. **g**, Quantification of γH2AX staining in cortical brain sections (*Casp1*$^{fl/fl}$ $n = 6$ and *Casp1*$^{fl/fl}$*Nes*^{cre} $n = 8$). Data are from two independent experiments. All *n* values refer to the number of mice used. Data are mean ± s.e.m. Statistics were calculated by unpaired two-tailed Student's *t*-test.

markedly enhanced staining of the DNA damage marker γH2AX in the brains of *Aim2*$^{-/-}$ mice (Fig. 3e–h). This increase in DNA damage was seen throughout the brain, including in regions that have been linked to fear and anxiety, such as the amygdala (Fig. 3g, h). We also examined levels of DNA damage in the adult cortex using single-cell gel electrophoresis (also known as the comet assay). These studies confirmed our previous γH2AX results and showed that deficits in AIM2 lead to substantially increased DNA damage accumulation, as indicated by the higher comet tail moment in the brains of AIM2-deficient mice (Fig. 3i, j). Similarly, we observed increased levels of γH2AX staining in the brains of *Ice*$^{-/-}$ and *Gsdmd*$^{-/-}$ mice (Fig. 3k). Although the nature and timing of these dsDNA breaks remain to be determined, it is evident that deficits in AIM2 inflammasome signalling result in increased accumulation of DNA damage in the brain. Collectively, these findings suggest that the AIM2 inflammasome and gasdermin-D aid in the removal of cells containing DNA damage from the brain.

CNS-derived inflammasomes shape behaviour

Mounting evidence suggests that immune activation in the periphery can have profound effects on brain maturation and behaviour[28]. Therefore, it is feasible that inflammasome signalling can shape behaviour and neurodevelopment both through its local actions in the brain and also via its functions in the periphery. To investigate this further, we first sought to identify which CNS-derived cell types express *Aim2* during neurodevelopment. Using fluorescence in situ hybridization, we found that *Aim2* is appreciably expressed by microglia, astrocytes and neurons in the developing brain (Extended Data Fig. 9a, b). Microglia are the innate immune sentinels of the brain and previous work suggests that microglia-coordinated innate immune responses can greatly affect brain development and function[29]. Notably, the deletion of caspase-1 in CX3CR1-expressing cells, which include microglia, did not result in the development of anxiety-related behaviours in either the open-field or the elevated plus maze tests (Extended Data Fig. 10a–c). By contrast, conditional ablation of caspase-1 from nestin-expressing CNS cells (that is, neurons, astrocytes and oligodendrocyte lineage cells) in *Casp1*$^{fl/fl}$*Nes*^{cre} mice led to anxiety-related behaviours and the accumulation of DNA damage in the brain (Fig. 4a–g). These data indicate a specific role for caspase-1 within the CNS in driving the observed behavioural phenotypes and preventing DNA damage accumulation.

Discussion

Our results underscore how deficits in the immune response to DNA insults can lead to impaired CNS development and neurological disease. The long-lived nature of neurons and glia, coupled with their exposure to high levels of replicative stress during neurodevelopment, makes the CNS especially vulnerable to DNA damage-induced dysfunction and pathology[18,30]. Yet, how the brain protects itself from genotoxic stress remains incompletely understood. Here we demonstrate that DNA damage surveillance by the AIM2 inflammasome is required for normal brain development and function. The AIM2 inflammasome and downstream gasdermin-D-mediated cell death contribute to the elimination of genetically compromised CNS cells. Furthermore, disruptions in this pathway lead to the development of anxiety-related behaviours, DNA damage accumulation in the CNS, and an increased number of neurons in the adult brain (Extended Data Fig. 10d). It is commonly assumed that CNS dieback solely occurs as a result of apoptotic cell death to remove unwanted cells. However, this assumption was made at a time when it was thought that there were only two forms of cell death—apoptosis and necrosis. Our findings demonstrating decreased CNS cell pruning and greater incorporation of neurons into the brains of gasdermin-D-deficient mice indicate that another form of cell death, namely pyroptosis, is involved in the sculpting of the brain. Further determination of the functional consequence of DNA damage sensing by the innate immune system may offer strategies for the treatment of a wide range of neurological disorders that are perpetuated by genotoxic stress.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2174-3.

1. Fernandes-Alnemri, T., Yu, J. W., Datta, P., Wu, J. & Alnemri, E. S. AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. *Nature* **458**, 509–513 (2009).
2. Hornung, V. et al. AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. *Nature* **458**, 514–518 (2009).
3. Rathinam, V. A. et al. The AIM2 inflammasome is essential for host defense against cytosolic bacteria and DNA viruses. *Nat. Immunol.* **11**, 395–402 (2010).
4. Guo, H., Callaway, J. B. & Ting, J. P. Inflammasomes: mechanism of action, role in disease, and therapeutics. *Nat. Med.* **21**, 677–687 (2015).
5. Choi, G. B. et al. The maternal interleukin-17a pathway in mice promotes autism-like phenotypes in offspring. *Science* **351**, 933–939 (2016).

6. Allan, S. M., Tyrrell, P. J. & Rothwell, N. J. Interleukin-1 and neuronal injury. *Nat. Rev. Immunol*. **5**, 629–640 (2005).
7. Felderhoff-Mueser, U., Schmidt, O. I., Oberholzer, A., Bührer, C. & Stahel, P. F. IL-18: a key player in neuroinflammation and neurodegeneration? *Trends Neurosci*. **28**, 487–493 (2005).
8. Dantzer, R., O'Connor, J. C., Freund, G. G., Johnson, R. W. & Kelley, K. W. From inflammation to sickness and depression: when the immune system subjugates the brain. *Nat. Rev. Neurosci*. **9**, 46–56 (2008).
9. Garber, C. et al. Astrocytes decrease adult neurogenesis during virus-induced memory dysfunction via IL-1. *Nat. Immunol*. **19**, 151–161 (2018).
10. Walsh, J. G., Muruve, D. A. & Power, C. Inflammasomes in the CNS. *Nat. Rev. Neurosci*. **15**, 84–97 (2014).
11. Yamaguchi, Y. & Miura, M. Programmed cell death in neurodevelopment. *Dev. Cell* **32**, 478–490 (2015).
12. Kuan, C. Y. et al. The Jnk1 and Jnk2 protein kinases are required for regional specific apoptosis during early brain development. *Neuron* **22**, 667–676 (1999).
13. Cecconi, F., Alvarez-Bolado, G., Meyer, B. I., Roth, K. A. & Gruss, P. Apaf1 (CED-4 homolog) regulates programmed cell death in mammalian development. *Cell* **94**, 727–737 (1998).
14. Yoshida, H. et al. Apaf1 is required for mitochondrial pathways of apoptosis and brain development. *Cell* **94**, 739–750 (1998).
15. Tzeng, T. C. et al. A fluorescent reporter mouse for inflammasome assembly demonstrates an important role for cell-bound and free ASC specks during in vivo infection. *Cell Rep*. **16**, 571–582 (2016).
16. Herzog, K. H., Chong, M. J., Kapsetaki, M., Morgan, J. I. & McKinnon, P. J. Requirement for Atm in ionizing radiation-induced cell death in the developing central nervous system. *Science* **280**, 1089–1091 (1998).
17. Takashima, H. et al. Mutation of *TDP1*, encoding a topoisomerase I-dependent DNA damage repair enzyme, in spinocerebellar ataxia with axonal neuropathy. *Nat. Genet*. **32**, 267–272 (2002).
18. McKinnon, P. J. Maintaining genome stability in the nervous system. *Nat. Neurosci*. **16**, 1523–1529 (2013).
19. McKinnon, P. J. Genome integrity and disease prevention in the nervous system. *Genes Dev*. **31**, 1180–1194 (2017).
20. Schumacher, B., Garinis, G. A. & Hoeijmakers, J. H. Age to survive: DNA damage and aging. *Trends Genet*. **24**, 77–85 (2008).
21. Hoeijmakers, J. H. DNA damage, aging, and cancer. *N. Engl. J. Med*. **361**, 1475–1485 (2009).
22. Hu, B. et al. The DNA-sensing AIM2 inflammasome controls radiation-induced cell death and tissue injury. *Science* **354**, 765–768 (2016).
23. Denes, A. et al. AIM2 and NLRC4 inflammasomes contribute with ASC to acute brain injury independently of NLRP3. *Proc. Natl Acad. Sci. USA* **112**, 4050–4055 (2015).
24. Adamczak, S. E. et al. Pyroptotic neuronal cell death mediated by the AIM2 inflammasome. *J. Cereb. Blood Flow Metab*. **34**, 621–629 (2014).
25. Wu, P. J., Liu, H. Y., Huang, T. N. & Hsueh, Y. P. AIM 2 inflammasomes regulate neuronal morphology and influence anxiety and memory in mice. *Sci. Rep*. **6**, 32405 (2016).
26. Kayagaki, N. et al. Caspase-11 cleaves gasdermin D for non-canonical inflammasome signalling. *Nature* **526**, 666–671 (2015).
27. Vanden Berghe, T., Linkermann, A., Jouan-Lanhouet, S., Walczak, H. & Vandenabeele, P. Regulated necrosis: the expanding network of non-apoptotic cell death pathways. *Nat. Rev. Mol. Cell Biol*. **15**, 135–147 (2014).
28. Filiano, A. J., Gadani, S. P. & Kipnis, J. How and why do T cells and their derived cytokines affect the injured and healthy brain? *Nat. Rev. Neurosci*. **18**, 375–384 (2017).
29. Hammond, T. R., Robinton, D. & Stevens, B. Microglia and the brain: complementary partners in development and disease. *Annu. Rev. Cell Dev. Biol*. **34**, 523–544 (2018).
30. Subba Rao, K. Mechanisms of disease: DNA repair defects and neurological disease. *Nat. Clin. Pract. Neurol*. **3**, 162–172 (2007).

Methods

Mice

All mouse experiments were performed in accordance with the relevant guidelines and regulations of the University of Virginia and approved by the University of Virginia Animal Care and Use Committee. Wild-type C57BL/6, *Aim2*$^{-/-}$[3], *Casp1*$^{-/-}$*Casp11*$^{-/-}$ (*Ice*$^{-/-}$)[31], *Nlrp3*$^{-/-}$[32], *Myd88*$^{-/-}$[33], *Il1r*$^{-/-}$[34], *Il18r*$^{-/-}$[35], R26-CAG-ASC-citrine[15], *Casp11*$^{-/-}$[36], *Nes*^{cre}[38], and *Cx3cr-1*^{cre}[39] mice were obtained from The Jackson Laboratory. *Gsdmd*$^{-/-}$ mice were provided by V. Dixit[26]. *Casp1*$^{fl/fl}$ mice were provided by R. Flavell[37]. Mice were housed and behaviour was conducted in specific pathogen-free conditions under standard 12 h light/dark cycle conditions in rooms equipped with control for temperature (21 ± 1.5 °C) and humidity (50 ± 10%).

Immunocytochemistry

Adult male mice were perfused with 4% paraformaldehyde in PBS. The brains were removed and fixed in 4% paraformaldehyde in PBS overnight at 4 °C. After dehydration in 30% sucrose, 30-µm sagittal sections were obtained using a Leica CM1950 cryostat (Leica). Sections were permeabilized with blocking solution containing 0.4% Triton X-100, 2% donkey serum, and 1% bovine serum albumin (BSA) in PBS for 1 h at room temperature and then incubated with primary antibodies overnight at 4 °C. Primary antibodies were diluted as follows: anti-GFAP (Invitrogen, 13-0300, 1:500), anti-calbindin (Sigma, C9848, 1:1,000), anti-γH2AX (abcam, ab11174, 1:1,000), anti-53BP1 (abcam, ab21083, 1:1,000). The next day, sections were incubated with fluorescently conjugated secondary antibodies (Invitrogen) for 2 h at room temperature, and mounted in ProLong Gold antifade reagent (Invitrogen). Images of stained brain sections were acquired using a confocal microscope (Leica TCS SP8) and analysed using ImageJ software. ASC visualization was accomplished using 488 nm detectors. Approximately two or three sections from each individual mouse were analysed and averages were used for data analysis.

TUNEL assay

Cell death was measured in vivo using a TUNEL assay (Roche, 11 684 795 910) according to manufacturer's instructions. In brief, brains from P5 mice were perfused with PBS followed by 4% PFA and then drop-fixed in 4% PFA for 24 h. After dehydration in 30% sucrose, 30-µm sagittal sections were obtained using a Leica CM1950 cryostat. Sections were permeabilized with blocking solution containing 0.4% Triton X-100, 2% donkey serum and 1% BSA in PBS for 1 h at room temperature and then incubated with primary antibodies overnight at 4 °C and secondary antibodies for 2 h at room temperature. Sections were mounted and allowed to dry on a slide before 50 µl of the TUNEL reaction mixture was added to each section. Slides were then incubated in a humidified atmosphere for 60 min at +37 °C in the dark. Slides were then rinsed three times with 1× PBS and then analysed under a fluorescence microscope using a 488-nm laser. Around two or three sections from each individual mouse were analysed and averages were used for data analysis.

Comet assay

The comet assay was performed using the Oxiselect Comet Assay Kit (Cell-Biolabs Inc.) according to the manufacturer's instructions with minor modifications. Slides were coated in low melting agarose the night before the assay and left to dry overnight. Cells were added to the top of agarose-coated slides followed by immediate placement of a cover slip to ensure an even and flat distribution of cells on the slide.

Mixed neuron/glia culture

Mixed CNS culture was performed according to a previously established protocol with minor modifications[40]. All cell culture plates were pre-coated with poly-D-lysine before seeding. P0 pups were euthanized and the brain was placed into cold Neuron-Glia dissection buffer (Minimum Essential Medium (MEM), Invitrogen, 11090). After removal of the meninges, the cortex was detached and placed in a 50-ml conical with cold Neuron-Glia dissection buffer. The cortices were triturated into a single cell suspension using serological pipette and 3×10^5 cells were plated per well in a 24-well plate. Cells were cultured for 10–14 days before treatment, with media changes every 2–3 days.

dsDNA stimulations

After 10–14 days in culture, cells were stimulated with or without LPS (0.5 µg ml^{-1}) in stimulation media containing Iscove's Modified Dulbecco's Medium (Gibco), 1% penicillin/strepomycin, 10% fetal bovine serum, 1% L-glutamine, and 50 µM 2-β-mercaptoethanol for 4 h at 37 °C. After 4 h, cells were given additional stimuli to induce/mimic dsDNA breaks.

Etoposide. A 100 mM stock of etoposide (abcam, ab120227) was prepared in DMSO according to the manufacturer's instructions and diluted to a final concentration of 20 µM or 100 µM in stimulation media. Cells were treated with or without etoposide and incubated overnight at 37 °C and supernatants were collected the next morning for additional assays.

Ionizing radiation. Cells were exposed to 40 Gy for 20 min and incubated overnight at 37 °C. Supernatants were collected the next morning for additional assays. P5 mice were exposed to 14 Gy of ionizing radiation and then returned to their home cage for 6 h. After 6 h, brain tissue was obtained for immunofluorescence and TUNEL staining.

Poly(dA:dT). Mixed glia cells were transfected with or without poly(dA:dT) using Lipofectamine 2000 reagent (Invitrogen, 11668-030) according to the manufacturer's instructions and incubated overnight at 37 °C. Supernatants were collected the following morning for additional assays.

Cytotoxicity

LDH release was measured using the CytoTox96 Non-Radioactive Cytotoxicity Assay (Promega, G1780) according to the manufacturer's instructions. Maximum LDH release control for each plate was generated using Lysis Solution.

RNA in situ hybridization

Brains from P5 pups were fixed in formalin for 48 h, embedded in paraffin, and cut into 5-µm sections. In situ hybridization was carried out according to the manufacturer's instructions (Affymetrix, QVT0012). Probes recognizing *Aim2* RNA (NM_001013779) were multiplexed with probes recognizing RNA expressed in neurons (*Rbfox3*; NM_001039167), microglia (*Aif1*; NM_019467), and astrocytes (*Gfap*; NM_010277). Ubiquitin (*Ubc*; NM_019639) was used as a positive control.

Behavioural testing

All behavioural testing was performed according to previously established behavioural methodologies. Behavioural experiments were carried out during daylight hours in a blinded fashion. All behaviour was carried out using adult male mice (8–12 weeks old).

Elevated plus maze

Anxiety was assessed using an elevated plus maze. The elevated plus maze consisted of two open arms (35×6 cm^2) and two closed arms (35×6 cm^2) with black plexiglass walls (20 cm in height) that extended from a common central platform (8×6 cm^2). The apparatus was constructed from polypropylene and Plexiglas (white floor, black walls) and elevated to a height of (121 cm) above floor level. Mice were individually placed on the centre square, facing an open arm, and allowed to freely explore the apparatus for 5 min. Activity was measured by a computer-assisted TopScan optical animal activity system (v.3.0).

Open-field testing

Spontaneous locomotor activity and anxiety was assessed in an open-field test. The open field consists of a square arena (40 × 40 cm^2) with white Plexiglas walls and floor, evenly illuminated. All mice were individually placed in the top left corner of the open field and left undisturbed to explore the arena over a 10 min session. Activity was measured by a computer-assisted TopScan optical animal activity system (v.3.0). Bouts into and time spent in a central square (15 × 15 cm^2) of the open field were automatically recorded as centre bouts and centre time, respectively. After the 10 min open-field exploration, mice were returned to their home cage and the number of urine stains and faecal pellets in the field were counted.

Sucrose preference

Depression-associated behaviours were assessed by measuring sucrose preference. Mice were given access to both untreated water and 2% sucrose water for 3 days. To prevent possible effects of side-preference in drinking behaviour, the position of the bottles in the cage was alternated every other day. No previous food or water deprivation was applied before the test.

Escape behaviour

Modified tail suspension test was used to measure escape behaviour. Mice were raised by their tail 30 cm into the air and assessed for escape behaviour. Mice were monitored until they became immobile.

Visual platform test

Visual performance was assessed using a visual platform in the Morris water maze. Mice were placed in clear water with a visible white platform and performance was evaluated as time spent reaching the platform.

Statistics and reproducibility

All statistical analyses were performed using GraphPad Prism. Statistical significance was calculated by unpaired two-tailed Student's *t*-test, one-way ANOVA with Tukey's post hoc tests or two-way ANOVA with Tukey's post hoc tests. $P < 0.05$ was considered significant. Sample size was chosen on the basis of similar previous studies[5,25], and not on statistical methods to predetermine sample size. All mouse experiments and data quantification were done in a blinded and randomized fashion. No statistical methods were used to predetermine sample size.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Source Data for Figs. 1–4 and Extended Data Figs. 1, 3–7, 9, 10 containing raw data for all experiments are provided with the paper. All other data are available from the corresponding author upon request.

31. Kuida, K. et al. Altered cytokine export and apoptosis in mice deficient in interleukin-1 beta converting enzyme. *Science* **267**, 2000–2003 (1995).
32. Kovarova, M. et al. NLRP1-dependent pyroptosis leads to acute lung injury and morbidity in mice. *J. Immunol.* **189**, 2006–2016 (2012).
33. Hou, B., Reizis, B. & DeFranco, A. L. Toll-like receptors activate innate and adaptive immunity by using dendritic cell-intrinsic and -extrinsic mechanisms. *Immunity* **29**, 272–282 (2008).
34. Glaccum, M. B. et al. Phenotypic and functional characterization of mice that lack the type I receptor for IL-1. *J. Immunol.* **159**, 3364–3371 (1997).
35. Hoshino, K. et al. Cutting edge: generation of IL-18 receptor-deficient mice: evidence for IL-1 receptor-related protein as an essential IL-18 binding receptor. *J. Immunol.* **162**, 5041–5044 (1999).
36. Wang, S. et al. Murine caspase-11, an ICE-interacting protease, is essential for the activation of ICE. *Cell* **92**, 501–509 (1998).
37. Case, C. L. et al. Caspase-11 stimulates rapid flagellin-independent pyroptosis in response to Legionella pneumophila. *Proc. Natl Acad. Sci. USA* **110**, 1851–1856 (2013).
38. Tronche, F. et al. Disruption of the glucocorticoid receptor gene in the nervous system results in reduced anxiety. *Nat. Genet.* **23**, 99–103 (1999).
39. Yona, S. et al. Fate mapping reveals origins and dynamics of monocytes and tissue macrophages under homeostasis. *Immunity* **38**, 79–91 (2013).
40. Chen, S. H., Oyarzabal, E. A. & Hong, J. S. Preparation of rodent primary cultures for neuron-glia, mixed glia, enriched microglia, and reconstituted cultures with microglia. *Methods Mol. Biol.* **1041**, 231–240 (2013).

Acknowledgements We thank members of the Lukens laboratory and the Center for Brain Immunology and Glia (BIG) for discussions. This work was supported by The Hartwell Foundation (Individual Biomedical Research Award to J.R.L.), a Rettsyndrome.org grant (22349 to J.R.L.), The Owens Family Foundation (awarded to J.R.L.), and a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (27515 to J.R.L.). C.R.L was supported by a NIH National Institute of General Medical Sciences predoctoral training grant (3T32GM008328) and a Wagner Fellowship. A.C.B. was supported by a Medical Scientist Training Program Grant (5T32GM007267-38) and an Immunology Training Grant (5T32AI007496-25). H.E.E. was supported by a Cell and Molecular Biology Training Grant (T32GM008136). E.L.F. was supported by a National Multiple Sclerosis Foundation Postdoctoral Fellowship (FG-1707-28590). C.E.B. was supported by Hutcheson and Stull Undergraduate Research Fellowships.

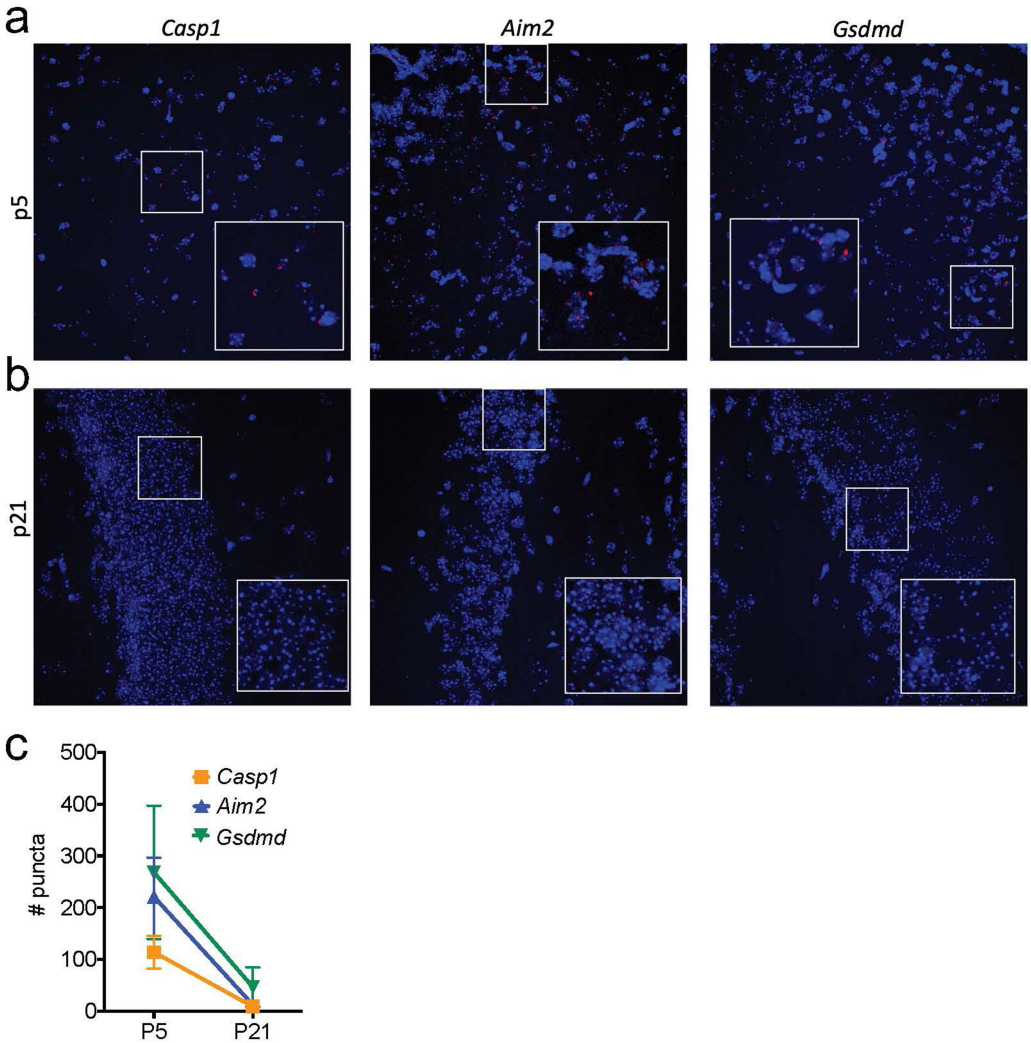
Author contributions C.R.L. and J.R.L. designed the study; C.R.L., E.L.F., C.E.B., A.C.B., C.A.M., M.E.H., M.J.P., H.E.E. and J.R.L. performed experiments; C.R.L. and J.R.L. analysed data and wrote the manuscript; J.R.L. oversaw the project.

Competing interests The authors declare no competing interests.

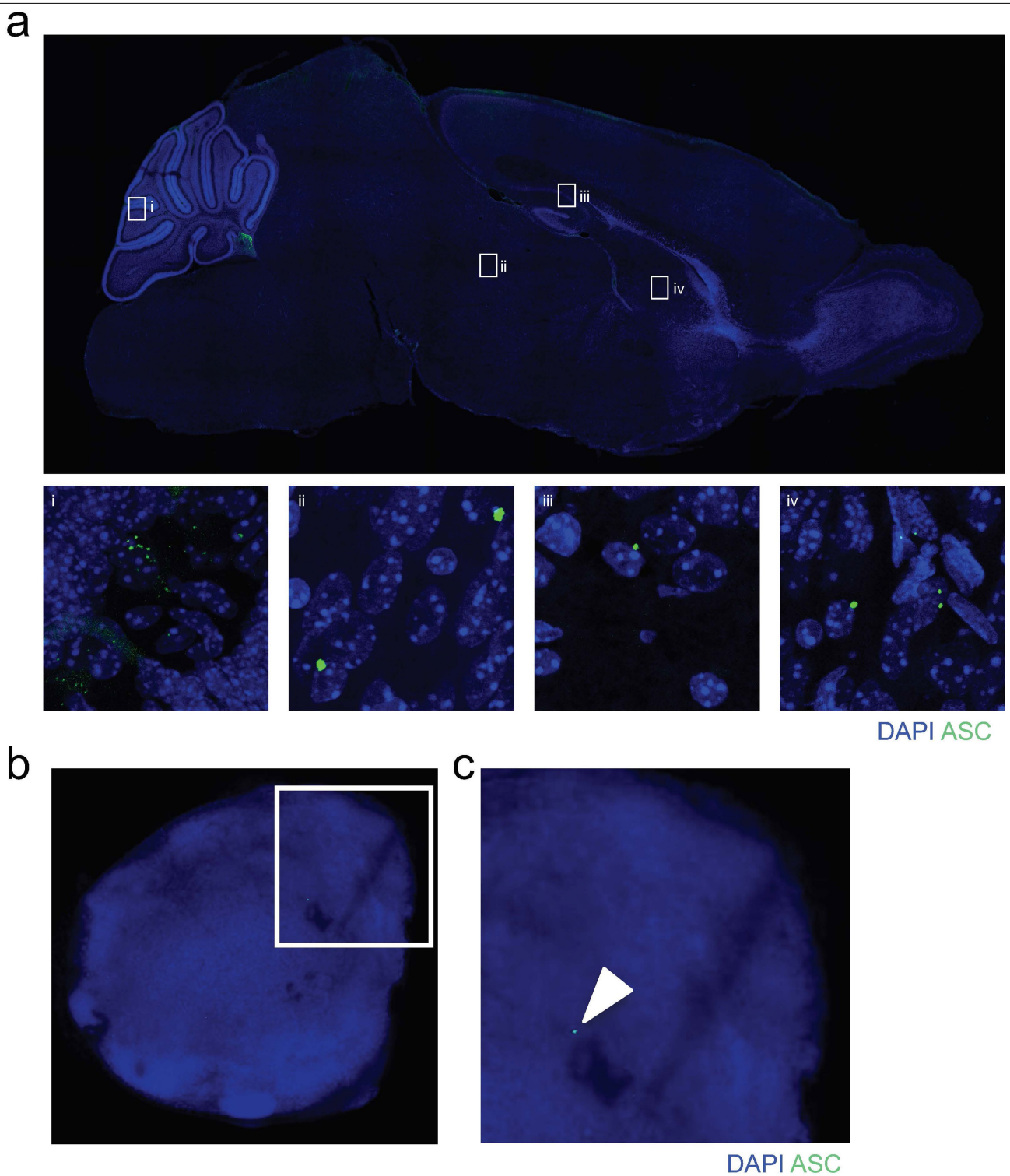
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2174-3.

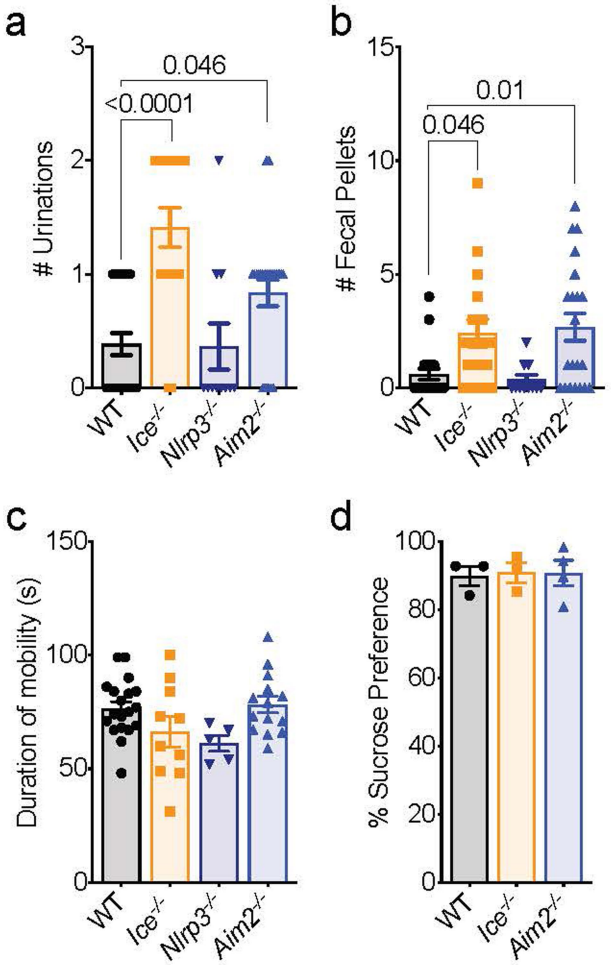
Correspondence and requests for materials should be addressed to J.R.L.



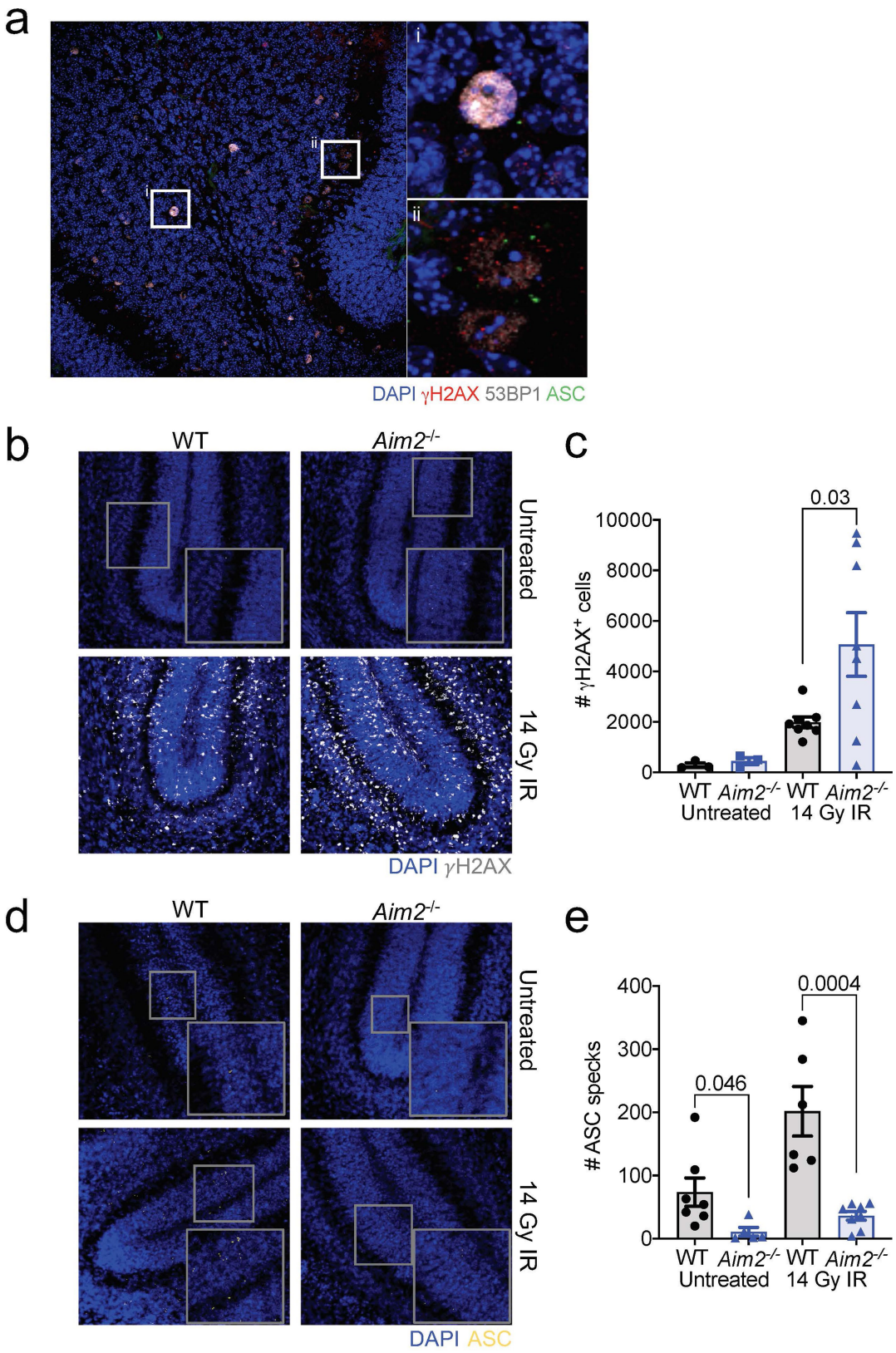
Extended Data Fig. 1 | Molecular components of the AIM2 inflammasome are abundantly expressed in the brain during neurodevelopment. a–**c**, Brains from P5 (**a**) and P21 (**b**) wild-type mice were evaluated for the mRNA expression of inflammasome component genes (red) *Casp1* (P5 $n = 4$, P21 $n = 2$), *Aim2* (P5 $n = 3$, P21 $n = 2$) and *Gsdmd* (P5 $n = 3$, P21 $n = 2$) using RNAscope. **c**, Quantification of *Casp1* (P5 $n = 4$, P21 $n = 2$), *Aim2* (P5 $n = 3$, P21 $n = 2$) and *Gsdmd* (P5 $n = 3$, P21 $n = 2$) mRNA puncta in the hippocampus per 40× image; from one experiment. *n* values refer to biological replicates. Data are mean ± s.e.m.



Extended Data Fig. 2 | ASC speck formation routinely occurs in the developing brain but is rare in mature lymph nodes under steady-state conditions. a, Top, sagittal image of ASC speck formation (green) in the brain of P5 ASC-citrine reporter mice. Original magnification, ×10. Bottom, ASC specks are detected throughout the brain using a ×40 objective including in the cerebellum (i), midbrain (ii), hippocampus (iii) and thalamus (iv). Representative images from three independent experiments with similar results. **b**, **c**, Adult (8–12 weeks old) ASC-citrine reporter mice were evaluated for peripheral inflammasome activation based on ASC speck formation (green) in the deep cervical lymph node (DCLN) using confocal microscopy with a ×10 objective (**b**). **c**, Arrow shows magnified image of ASC speck (green) formed in the DCLN. Representative images from two independent experiments with similar results.

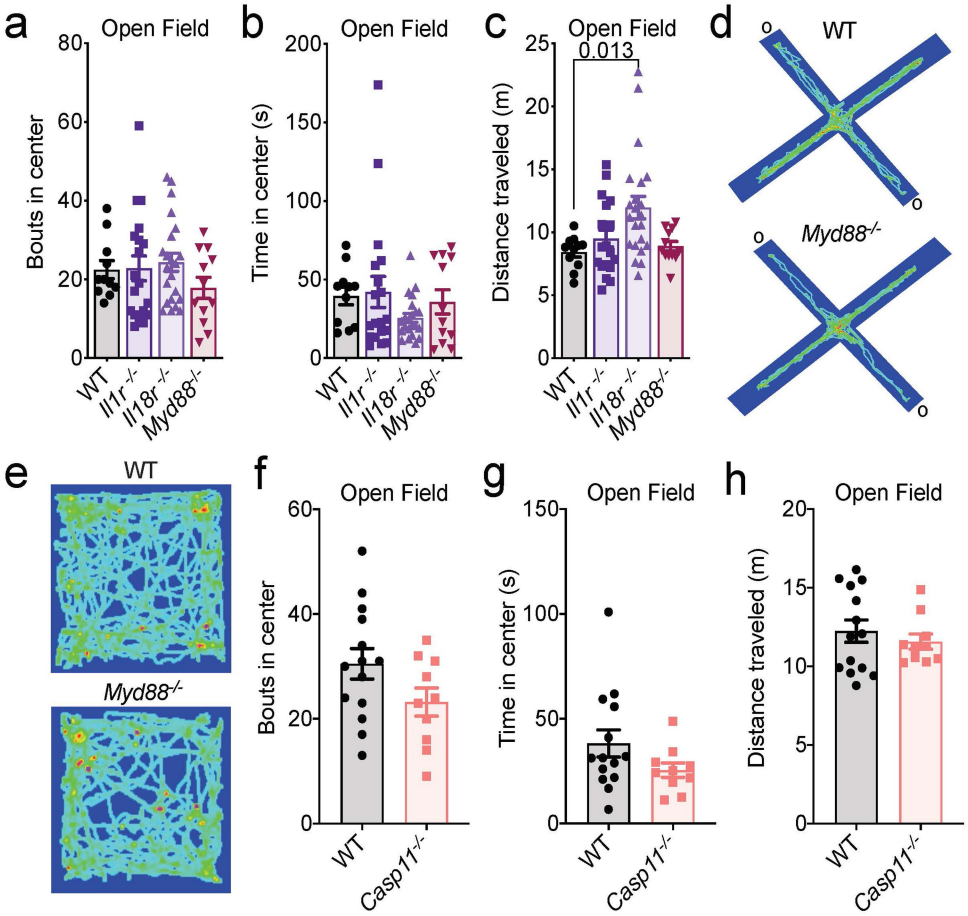


Extended Data Fig. 3 | Lack of AIM2 inflammasome signalling results in an increase in anxiety-related behaviours but not depressive-related behaviours. Adult (8–12 weeks old) wild-type, *Ice*$^{-/-}$, *Nlrp3*$^{-/-}$ and *Aim2*$^{-/-}$ mice were assessed for behavioural abnormalities. **a**, **b**, Number of urinations (WT $n = 26$, *Ice*$^{-/-}$ $n = 17$, *Nlrp3*$^{-/-}$ $n = 11$, *Aim2*$^{-/-}$ $n = 24$; from three independent experiments) (**a**) and number of faecal pellets (WT $n = 20$, *Ice*$^{-/-}$ $n = 17$, *Nlrp3*$^{-/-}$ $n = 11$, *Aim2*$^{-/-}$ $n = 21$; from three independent experiments) (**b**) were measured during 10 min of open-field testing. **c**, **d**, Depressive behaviours were evaluated in adult male mice using the tail suspension test for escape behaviour (WT $n = 19$, *Ice*$^{-/-}$ $n = 10$, *Nlrp3*$^{-/-}$ $n = 5$, *Aim2*$^{-/-}$ $n = 14$; from two independent experiments) (**c**) and sucrose preference test (WT $n = 3$, *Ice*$^{-/-}$ $n = 3$, *Aim2*$^{-/-}$ $n = 4$; from one independent experiment) (**d**). All *n* values refer to the number of mice used. Data are mean ± s.e.m. *P* values were determined by one-way ANOVA with Tukey's post hoc tests.

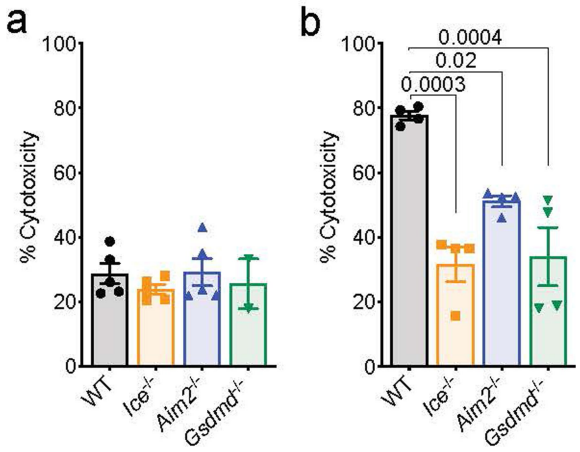


Extended Data Fig. 4 | See next page for caption.

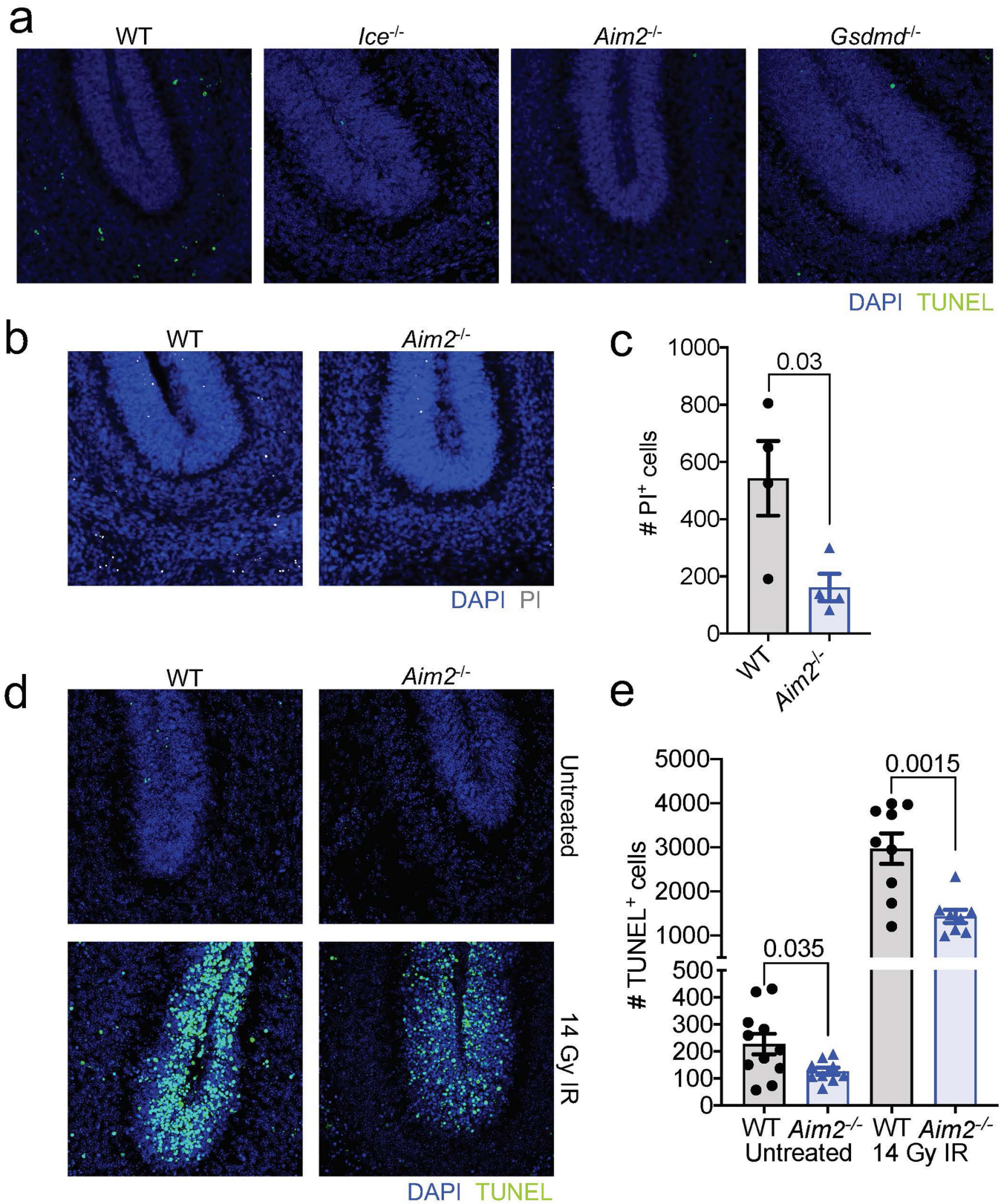
Extended Data Fig. 4 | ASC specks form in response to DNA damage in the developing brain. a, Brains from wild-type P5 mice were evaluated for localization of ASC specks (green) in relation to $DAPI^+$ nuclei (blue) containing DNA damage (γH2AX (red), 53BP1 (grey)) in the cerebellum. **i**, **ii**, Magnified images in regions showing ASC specks formed in close proximity to nuclei containing DNA damage. Representative images from four mice with similar results, from one experiment. Original magnifications, ×40. Differences in the size of nuclei probably reflect specific stages in replication, DNA repair, differentiation, or cell death that the individual cells are in, as well as differences seen across CNS cell types. **b**–**e**, Wild-type and *Aim2*$^{-/-}$ P5 mice received either control treatment or 14 Gy of ionizing radiation (IR) to induce DNA damage. Brains were obtained 6 h later, and immunostained to measure DNA damage induction (γH2AX staining) and inflammasome activation (ASC speck formation) in the cerebellum. **b**, Representative cerebellar images of γH2AX staining; from two independent experiments with similar results. Original magnification, ×20. **c**, Quantification of γH2AX staining in the cerebellum (untreated: WT $n = 3$, *Aim2*$^{-/-}$ $n = 3$; IR-treated: WT $n = 8$, *Aim2*$^{-/-}$ $n = 8$; from two independent experiments). **d**, Representative cerebellar images of ASC speck formation; from two independent experiments with similar results. Original magnification, ×20. **e**, Quantification of ASC speck formation in the cerebellum (untreated: WT $n = 7$, *Aim2*$^{-/-}$ $n = 5$; IR-treated: WT $n = 6$, *Aim2*$^{-/-}$ $n = 8$; from two independent experiments). All n values refer to the number of mice used. Data are mean ± s.e.m. P values were determined by unpaired two-tailed Student's t-test.



Extended Data Fig. 5 | Anxiety phenotypes do not develop in mice that lack IL-1R, IL-18R, MYD88 or caspase-11. All behavioural testing was conducted on adult (8–12 weeks old) mice. **a**–**c**, Behaviours for anxiety were evaluated by bouts into (**a**) and time spent in (**b**) the centre of the open-field arena, and the total distance travelled (**c**) (WT $n = 11$, *Il1r*$^{-/-}$ $n = 19$, *Il18r*$^{-/-}$ $n = 22$, *Myd88*$^{-/-}$ $n = 12$; from three independent experiments). **d**, Anxiety-related behaviours were assessed in wild-type and *Myd88*$^{-/-}$ mice using the elevated plus maze. Representative heat maps from four independent experiments with similar results depicting path of travel through open and closed arms of the maze. **e**, Representative heat maps from four independent experiments with similar results of the path travelled by adult wild-type and *Myd88*$^{-/-}$ mice in the open-field arena. **f**–**h**, Quantification of bouts into (**f**) and time spent in (**g**) the centre of the open-field arena, and the total distance travelled (**h**) (WT $n = 14$, *Casp11*$^{-/-}$ $n = 10$; from two independent experiments). All n values refer to the number of mice used. Data are mean ± s.e.m. No statistically significant differences were determined by one-way ANOVA with Tukey's post hoc tests (**a**–**c**) or unpaired two-tailed Student's t-test (**f**–**h**).

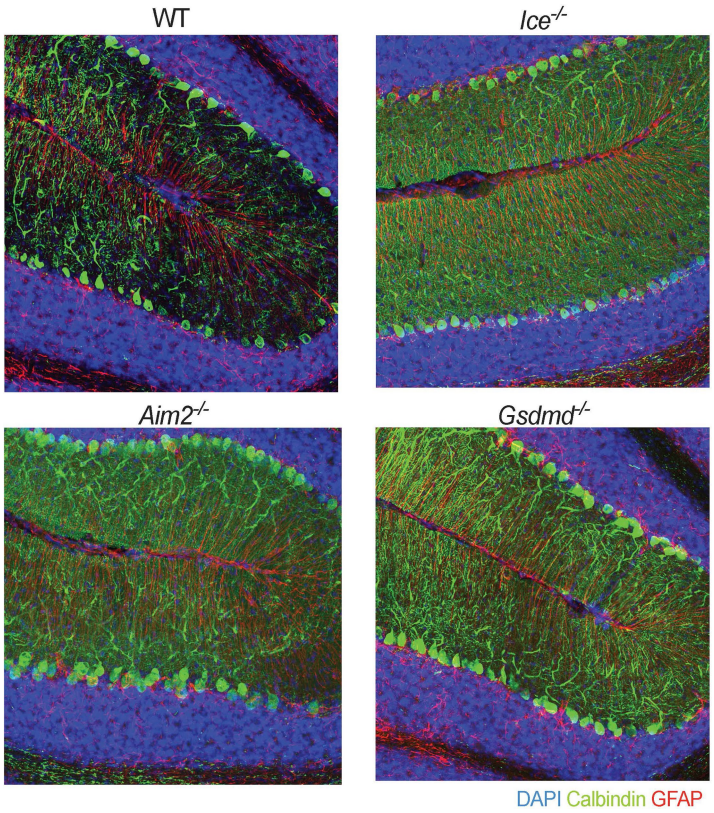


Extended Data Fig. 6 | Genetic ablation of the AIM2 inflammasome or gasdermin-D in CNS cells limits cell death in response to DNA insults. Mixed neural cultures were generated from wild-type, *Ice*$^{-/-}$, *Aim2*$^{-/-}$ and *Gsdmd*$^{-/-}$ P0 mice. **a**, Mixed neural cell cultures were left untreated to test for baseline differences in cytotoxicity (WT $n = 5$, *Ice*$^{-/-}$ $n = 5$, *Aim2*$^{-/-}$ $n = 5$, *Gsdmd*$^{-/-}$ $n = 2$). **b**, Mixed neural cell cultures were primed with LPS for 4 h followed by transfection with poly(dA:dT) (WT $n = 4$, *Ice*$^{-/-}$ $n = 4$, *Aim2*$^{-/-}$ $n = 4$, *Gsdmd*$^{-/-}$ $n = 4$). Cell death was measured by LDH release after overnight stimulation. Representative data from three independent experiments. All *n* values refer to biological replicates from one representative experiment. Data are mean ± s.e.m. *P* values were determined by one-way ANOVA with Tukey's post hoc tests.

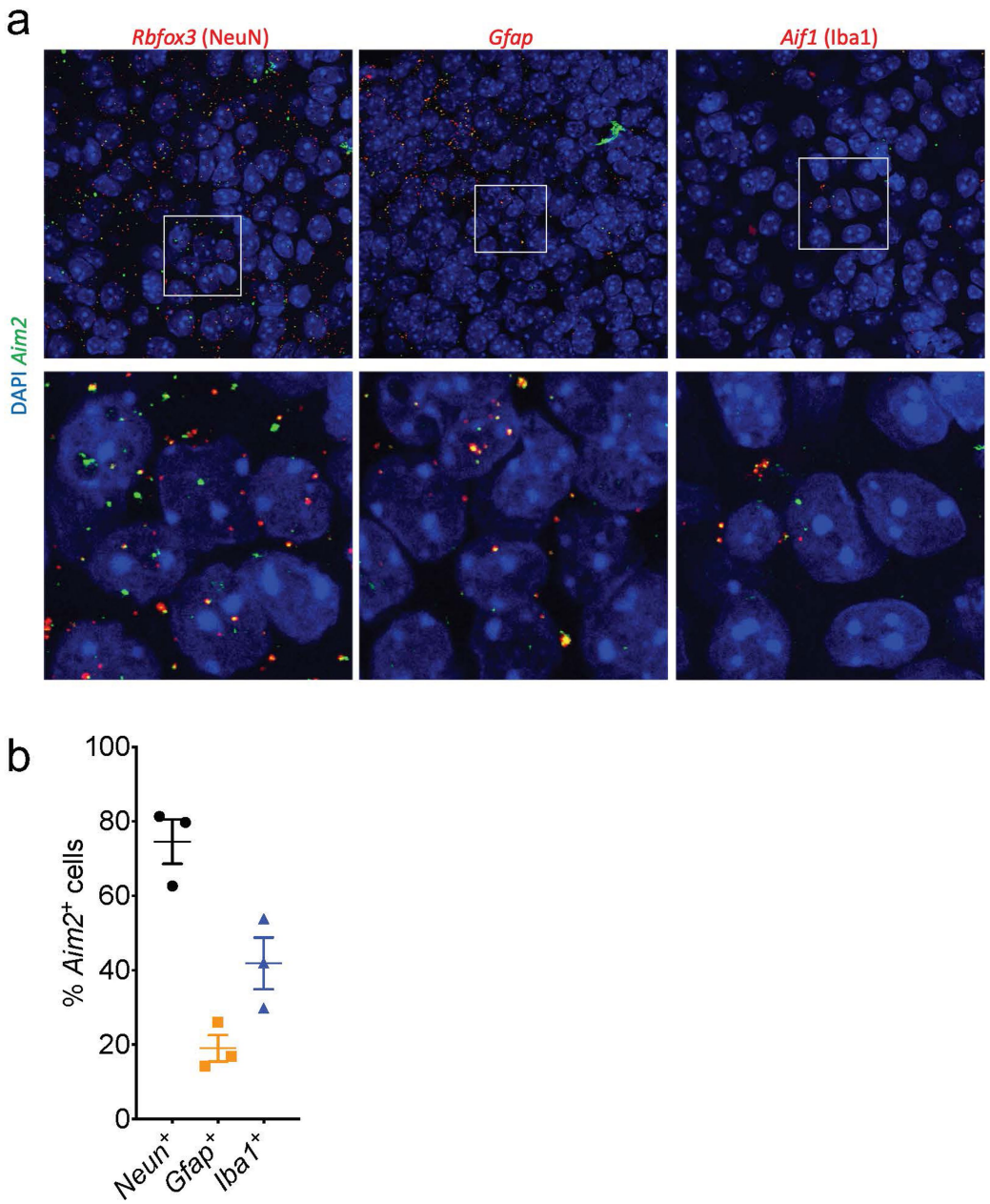


Extended Data Fig. 7 | See next page for caption.

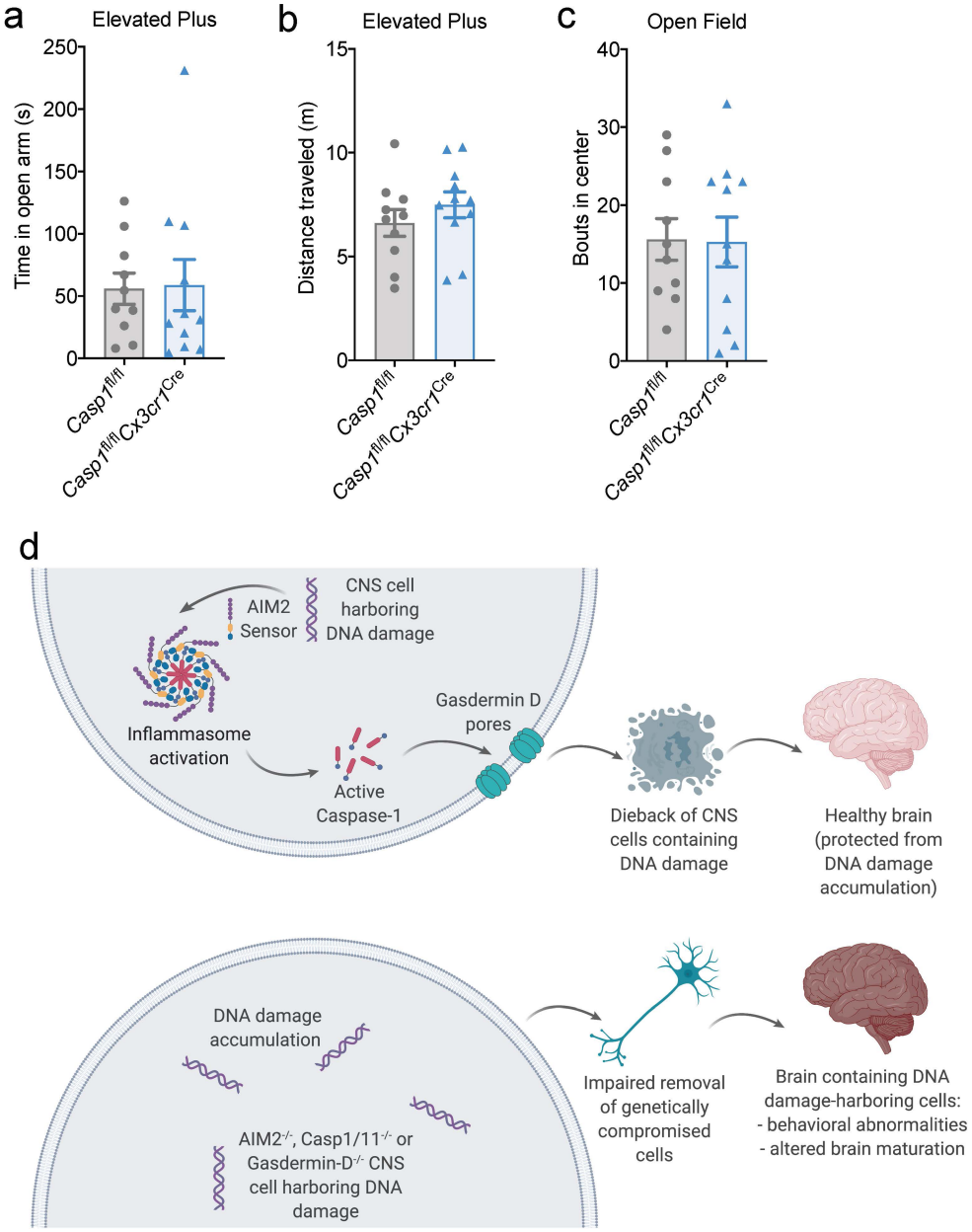
Extended Data Fig. 7 | AIM2 contributes to CNS cell death during neurodevelopment and in response to ionizing radiation. a, Representative images from wild-type, *Ice*$^{-/-}$, *Aim2*$^{-/-}$ and *Gsdmd*$^{-/-}$ P5 mice showing TUNEL^{+} cells (green) in the cerebellum. Images are representative of two independent experiments with similar results. Original magnifications, ×20. **b**, Representative images of additional markers of cell death (propidium iodide (PI), grey) in wild-type and *Aim2*$^{-/-}$ P5 mice. Original magnifications, ×20. **c**, Quantification of propidium iodide-positive cells in the cerebellum of WT ($n = 4$) and *Aim2*$^{-/-}$ ($n = 4$) P5 mice; from one independent experiment. **d**, **e**, Wild-type and *Aim2*$^{-/-}$ P5 mice received either control treatment or 14 Gy of ionizing radiation to induce DNA damage. Brains were obtained 6 h later and the TUNEL assay was conducted on cerebellar sections to evaluate cell death. **d**, Representative images showing TUNEL staining in the cerebellum of untreated and irradiated wild-type and *Aim2*$^{-/-}$ P5 mice; from three independent experiments with similar results. Original magnifications, ×20. **e**, Quantification of TUNEL^{+} cells in the cerebellums of untreated and irradiated wild-type ($n = 11$ untreated, $n = 9$ IR) and *Aim2*$^{-/-}$ ($n = 9$ untreated, $n = 8$ IR) mice; from three independent experiments. All n values refer to the number of mice used. Data are mean ± s.e.m. P values were determined by unpaired two-tailed Student's t-test.



Extended Data Fig. 8 | Lack of AIM2 inflammasome components increases the number of Purkinje neurons that are incorporated into the adult brain. Representative images of cerebellums from adult (8–12 weeks old) wild-type, *Ice*$^{-/-}$, *Aim2*$^{-/-}$ and *Gsdmd*$^{-/-}$ mice showing an increase in the number of Purkinje cells (calbindin^{+} cells) in mice lacking inflammasome components. Original magnifications, ×20. Images are representative of three independent experiments with similar results.



Extended Data Fig. 9 | *Aim2* is expressed by neurons, astrocytes and microglia in the developing brain. Brains from wild-type P5 mice ($n = 3$; from 1 experiment) were evaluated for expression of *Aim2* using RNAscope. **a**, Images showing co-expression of *Aim2* (green) and CNS cell-specific genes *Rbfox3*: NeuN (red), *Gfap*: GFAP (red), and *Aif1*: Iba1 (red) in the hippocampus. Original magnifications, ×40. **b**, Quantification of the percentage of CNS cells in images that are positive for *Aim2*. *n* values refer to biological replicates. Data are mean ± s.e.m.



Extended Data Fig. 10 | Deletion of caspase-1 in CX3CR1-expressing cells does not result in the development of anxiety-related behaviours. a–**c**, Adult (8–12 weeks old) *Casp1*$^{fl/fl}$ ($n = 10$) and *Casp1*$^{fl/fl}$*Cx3cr1*^{cre} ($n = 11$) mice were evaluated for anxiety-related behaviours using the time spent in the open arms (**a**) and distance travelled (**b**) in the elevated plus maze, along with total bouts into the centre of the open-field arena (**c**). Data are from two independent experiments. All *n* values refer to the number of mice used. **d**, Schematic of the proposed role that DNA damage surveillance by the AIM2 inflammasome has in neurodevelopment. Data are mean ± s.e.m. *P* values determined by unpaired two-tailed Student's *t*-test show no statistically significant differences. Graphical illustrations were made using BioRender (https://biorender.com/).

nature research

Corresponding author(s): Lukens, John R

Last updated by author(s): Jan 31, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Top Scan Software version 3.00 was used to track and measure all behaviors.
Data analysis	GraphPad Prism 8 was used to represent data in graphs and for the statistical analyses of the data. The ImageJ software was used to outline and measure areas for imaging analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Source Data for Figs. 1–4 and Extended Data Figs. 1, 3–7, and 9-10 containing raw data for all experiments, are provided with the paper. All other data are available from the corresponding author upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was chosen based on previous experiments and publications (5) (25), PMID:31482844, PMID:26311765.
Data exclusions	No data were excluded.
Replication	Key experiments were reiterated with representatives from each group with similar observations across iterations
Randomization	Subject animals were randomly assigned to experimental groups when applicable. Groups were determined based on genotype
Blinding	Investigators were blinded as to experimental groups during data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	rat anti-GFAP (Invitrogen, 13-0300, Lot SA247423 1:500), mouse anti-calbindin (Sigma, C9848, Lot 058M4813V 1:1,000), rabbit anti-γH2AX (abcam, ab11174, Lot GR3280493-2 1:1,000), rabbit anti-53bp1 (abcam, ab21083, Lot GR3266130-1 1:1,000), Biotin conjugated NeuN (Millipore, MAB377B, Lot 3232375, 1:300)
Validation	These are all commonly used antibodies. Statements regarding validation can be found at the manufactures website.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Wild-type (WT) C57BL/6, Aim2-/-3, Casp1/11-/- (Ice-/-)33, Nlrp3-/-34, Myd88-/-35, Il1r-/-36, Il18r-/-37, R26-CAG-ASC-citrine5, Casp11-/-38, Casp1fl/fl39, NestinCre40, and Cx3cr1Cre41 mice were obtained from The Jackson Laboratory. Gsdmd-/- mice were generously provided by Vishva Dixit42. Mice were housed and behavior was conducted in specific pathogen-free conditions under standard 12hr light/dark cycle conditions in rooms equipped with control for temperature (21 ± 1.5oC) and humidity (50 ± 10%). For behavior, adult male (8-12 week old) were used, for postnatal day 5 (p5) experiments a mix of male and female mice were used.
Wild animals	Wild animals were not used in our studies.
Field-collected samples	Field-collected samples were not used in our studies.
Ethics oversight	All mouse experiments were performed in accordance with the relevant guidelines and regulations of the University of Virginia and approved by the University of Virginia Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

A plant genetic network for preventing dysbiosis in the phyllosphere

https://doi.org/10.1038/s41586-020-2185-0

Received: 1 September 2019

Accepted: 19 February 2020

Published online: 8 April 2020

Check for updates

Tao Chen[1,2,3,8], Kinya Nomura[1,8], Xiaolin Wang[4], Reza Sohrabi[1,5], Jin Xu[6], Lingya Yao[4], Bradley C. Paasch[1], Li Ma[1], James Kremer[1], Yuti Cheng[1,3], Li Zhang[1,3], Nian Wang[6], Ertao Wang[4], Xiu-Fang Xin[4,7]✉ & Sheng Yang He[1,3,5]✉

The aboveground parts of terrestrial plants, collectively called the phyllosphere, have a key role in the global balance of atmospheric carbon dioxide and oxygen. The phyllosphere represents one of the most abundant habitats for microbiota colonization. Whether and how plants control phyllosphere microbiota to ensure plant health is not well understood. Here we show that the *Arabidopsis* quadruple mutant (*min7 fls2 efr cerk1*; hereafter, *mfec*)[1], simultaneously defective in pattern-triggered immunity and the MIN7 vesicle-trafficking pathway, or a *constitutively activated cell death1* (*cad1*) mutant, carrying a S205F mutation in a membrane-attack-complex/perforin (MACPF)-domain protein, harbour altered endophytic phyllosphere microbiota and display leaf-tissue damage associated with dysbiosis. The Shannon diversity index and the relative abundance of Firmicutes were markedly reduced, whereas Proteobacteria were enriched in the *mfec* and *cad1*^{S205F} mutants, bearing cross-kingdom resemblance to some aspects of the dysbiosis that occurs in human inflammatory bowel disease. Bacterial community transplantation experiments demonstrated a causal role of a properly assembled leaf bacterial community in phyllosphere health. Pattern-triggered immune signalling, MIN7 and CAD1 are found in major land plant lineages and are probably key components of a genetic network through which terrestrial plants control the level and nurture the diversity of endophytic phyllosphere microbiota for survival and health in a microorganism-rich environment.

The phyllosphere is inhabited by a diverse microbiota, with some phyllosphere microorganisms living on the surface of plants as epiphytes and others residing inside leaves as endophytes[2,3]. In contrast to the intensively studied roles of root-colonizing microbiota in plant health[4–11], the collective community-level contribution of phyllosphere microbiota to plant growth, development and health is not well understood. The phyllosphere is functionally distinct from the belowground rhizosphere. For example, compared with roots, leaves have a larger apoplast, which facilitates the gas exchange essential for photosynthesis and provides a largely air-filled internal space for microbiota colonization. The composition of leaf microbiota can be influenced by host genotypes[12–14], and a recent ecological study showed a positive correlation between leaf bacterial diversity and terrestrial ecosystem productivity[15]. However, whether these variations in phyllosphere microbiota make a causal contribution to (or are merely a consequence of) plant health remains an unresolved fundamental question.

In a previous study aimed at identifying plant pathways that are attacked by the bacterial pathogen *Pseudomonas syringae* pv. tomato (*Pst*) DC3000, we identified *Arabidopsis* quadruple mutants (for example, *mfec* and *min7 bak1 bkk1 cerk1* (hereafter *mbbc*))[1] that allowed increased proliferation of a nonpathogenic mutant of *Pst* DC3000 and harboured a greater abundance of endophytic leaf microorganisms under high humidity, one of the most common environmental conditions plants encounter in nature. *mfec* and *mbbc* mutants are defective in two pathways—pattern-triggered immune signalling[16,17] and the MIN7 vesicle-trafficking pathway, which is involved in modulating an aqueous microenvironment in the apoplast[1]. The *mfec* and *mbbc* mutants also exhibited spontaneous leaf necrosis and chlorosis in the absence of pathogen inoculation under high humidity[1]. However, the lack of well-controlled soil-based gnotobiotic plant growth systems (the *mfec* and *mbbc* mutants showed most obvious phenotypes when grown in soil) prevented us from answering the fundamental question of whether altered leaf endophytic microbiota are the cause or a consequence of poor phyllosphere health of *mfec* and *mbbc* plants.

Genotype-dependent shift of microbiota

Because *mfec* and *mbbc* exhibited similar phenotypes in initial experiments, we conducted further characterization using the *mfec* mutant and observed additional phenotypes. First, chlorosis and/or necrosis

[1]Department of Energy Plant Research Laboratory, Michigan State University, East Lansing, MI, USA. [2]State Key Laboratory of Agriculture Microbiology, Huazhong Agricultural University, Wuhan, China. [3]Howard Hughes Medical Institute, Michigan State University, East Lansing, MI, USA. [4]National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China. [5]Plant Resilience Institute, Michigan State University, East Lansing, MI, USA. [6]Citrus Research and Education Center, Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences, University of Florida, Lake Alfred, FL, USA. [7]CAS-JIC Center of Excellence for Plant and Microbial Sciences (CEPAMS), Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China. [8]These authors contributed equally: Tao Chen, Kinya Nomura. ✉e-mail: xinxf@sippe.ac.cn; hes@msu.edu

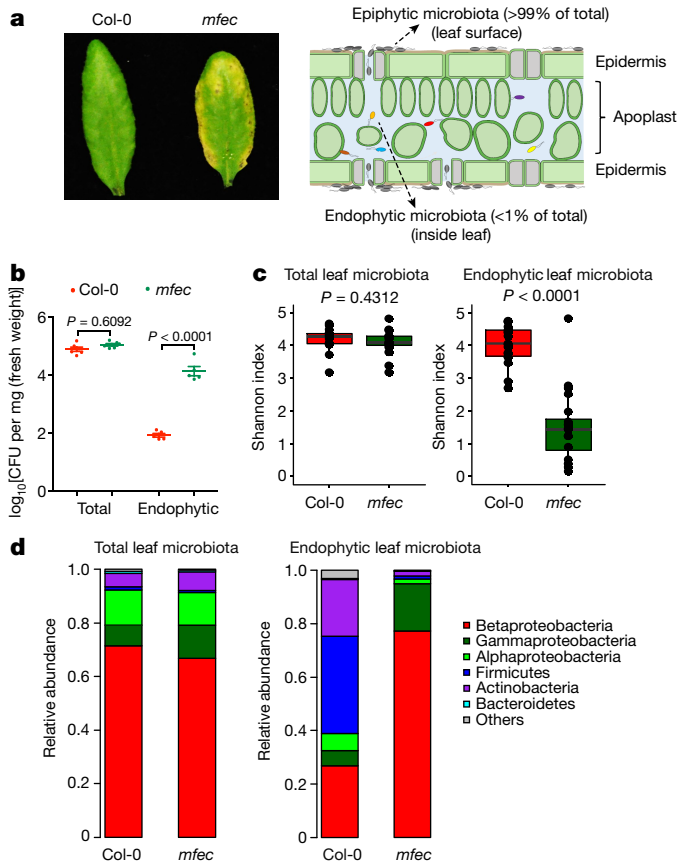


Fig. 1 | Total and endophytic leaf microbiota in Col-0 and *mfec* plants. **a**, **b**, Leaf appearance (**a**) and population size of leaf microbiota (**b**) in five-week-old Col-0 and *mfec* plants grown in *Arabidopsis* mix potting soil. **a**, Left, images at day 5 after plants (grown at approximately 60% humidity) were exposed to approximately 95% humidity. **a**, Right, cartoon showing epiphytic and endophytic microbiota in a leaf cross-section. Experiments in **a**, **b**, were repeated three times with similar results. One-way ANOVA with Tukey's test. $n = 6$ (total bacteria populations) and $n = 5$ (endophytic bacteria populations) biological replicates. Data are mean ± s.e.m. **c**, **d**, Shannon indexes (**c**) and the relative abundance of bacteria at the phylum level (**d**), obtained from 16S rRNA gene-sequence profiles of total and endophytic bacteria in Col-0 and *mfec* plants grown in *Arabidopsis* mix potting soil. In box plots, the centre line represents the median, box edges show the 75th and 25th percentiles, and whiskers extend to 1.5× the interquartile range. Two-tailed Mann–Whitney *U*-test. $n = 15$ (Col-0) and $n = 15$ (*mfec*) biological replicates for analysis of total leaf bacterial microbiota across 3 independent experiments; $n = 18$ (Col-0) and $n = 20$ (*mfec*) biological replicates for analysis of leaf endophytic bacterial microbiota across 4 independent experiments.

phenotypes in the *mfec* mutant were seen in plants grown in different soil types in air-circulating growth chambers, albeit to varying degrees (Fig. 1a, Extended Data Fig. 1a). Second, tissue damage appeared to be restricted mostly to leaves, as roots of Col-0 and the *mfec* mutant appeared similar (Extended Data Fig. 1b). Third, in contrast to the marked difference in the levels of endophytic bacteria (estimated after surface sterilization to remove epiphytic bacteria), little difference was observed between Col-0 and *mfec* plants in total leaf bacteria (without surface sterilization), which include both epiphytic and endophytic bacteria (Fig. 1b). The total bacteria count was usually at least 100-fold higher than that of endophytic bacteria (Fig. 1b; in Col-0). Our results provided evidence for compartment-specific modulation of phyllosphere bacteria and suggest that the bulk epiphytic phyllosphere bacteria may have a less intimate interaction with and, therefore, may be less influenced by host innate immune signalling and the MIN7 vesicle-trafficking pathway.

Next, we conducted 16S rRNA gene-sequencing analysis of leaf bacterial communities in Col-0 and *mfec* plants. Plants were grown in '*Arabidopsis* mix' potting soil in air-circulating growth chambers for colonization of phyllosphere microbiota. We observed that the endophytic leaf community in Col-0 plants was substantially more diverse than that in the *mfec* plants, as judged by Shannon index and observed operational taxonomic units (OTUs) (Fig. 1c, Extended Data Fig. 2a, b). Furthermore, Firmicutes and Actinobacteria were abundantly observed in Col-0 leaves, whereas their relative abundance was greatly reduced in the *mfec* mutant. Conversely, Betaproteobacteria and Gammaproteobacteria were highly enriched in *mfec* leaves (Fig. 1d). By contrast, there was no significant difference in bacterial composition of total leaf bacteria between Col-0 and *mfec* (Fig. 1c, d), providing further evidence for profound compartment-specific modulation of the level and composition of endophytic leaf microbiota by pattern-triggered immunity and the MIN7 vesicle-trafficking pathway.

To determine the individual contributions of pattern-triggered immunity and the MIN7 vesicle-trafficking pathway to the endophytic leaf microbiota shift in the *mfec* mutant, we performed further 16S rRNA gene sequencing. In these experiments, we inoculated the *Arabidopsis* mix potting soil with a 48-member leaf endophytic bacterial community derived from healthy Col-0 plants (SynCom$^{Col\text{-}0}$; Supplementary Table 1) as a consistent microbiota source, in addition to presumably variable soil and/or air-derived communities to which plants were exposed in growth chambers. Again, we observed a significant reduction in the overall diversity and a substantial shift in the composition of the endophytic bacterial community in the *mfec* leaves, but not in *fec* (defective in pattern-triggered immunity alone) or *min7* (defective in MIN7 pathway alone) leaves, compared with that in Col-0 leaves (Fig. 2a, b, Extended Data Fig. 2c). Correspondingly, only *mfec* quadruple-mutant plants displayed necrosis and chlorosis and had a higher level of endophytic bacterial microbiota (Fig. 2c, d). These results show non-redundant and essential roles of pattern-triggered immunity and the MIN7 vesicle-trafficking pathway in controlling the endophytic leaf microbiota in *Arabidopsis*.

Further analysis of endophytic microbiota 16S-profiling data from SynCom$^{Col\text{-}0}$-supplemented experiments revealed changes in specific amplicon sequence variants (ASV) representing distinct bacterial 16S rRNA gene sequences between Col-0 and *mfec* (Supplementary Table 2). ASVs belonging to Comamonadaceae (ASV1, ASV113, ASV141 and ASV280), Xanthomonadaceae (ASV12 and ASV386), Alcaligenaceae (ASV3) and Sphingomonadaceae (ASV245) were enriched in *mfec* plants (together representing 91.97% of reads in *mfec* plants compared with only 39.69% in Col-0 plants). Conversely, 33 Paenibacillaceae ASVs were depleted in *mfec* plants (representing 24.83% and 0.52% of reads in Col-0 and *mfec* plants, respectively). However, the observed modest enrichment of the ASVs—on the basis of relative abundance—may not account for the approximately 100-fold increase of the total endophytic microbiome population in the *mfec* mutant, suggesting that other ASVs could contribute to the increase of the total load of endophytic microbiota in *mfec* leaves without being reflected in their relative abundance. A clear resolution of all ASVs will require a further study using methods that are more appropriate for estimating the absolute abundance of ASVs.

Role of *mfec* microbiota in dysbiosis

The reduction of the overall relative bacterial diversity and conversion of a Firmicutes-rich community to a Proteobacteria-rich community in the *mfec* mutant was intriguing because these changes bear some resemblance to those observed in human microbiome dysbiosis associated with inflammatory bowel disease[18,19]. This raised the possibility that tissue damage in *mfec* plants may result from a form of dysbiosis in plants. However, true dysbiosis implies a causative role of altered microbiota in inducing symptoms. To test this possibility, we grew plants in sterile 0.5× Murashige–Skoog agar plates and in a peat-based gnotobiotic

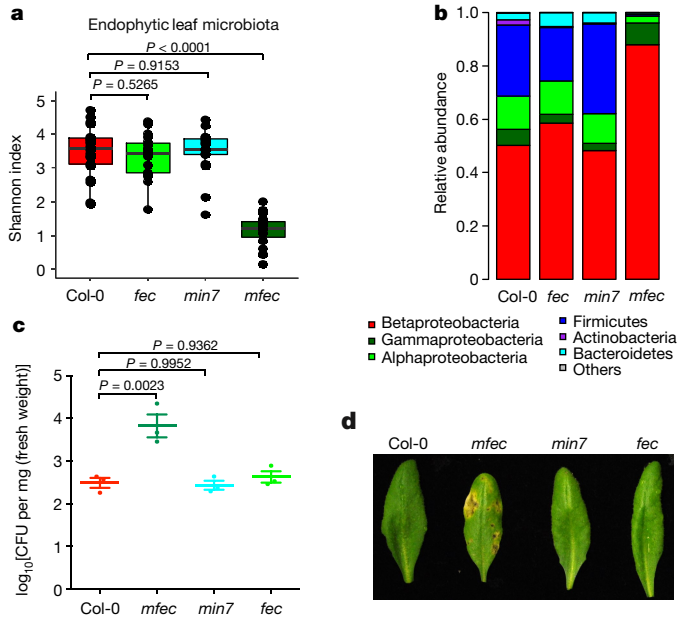


Fig. 2 | Endophytic leaf microbiota in Col-0, *fec*, *min7* and *mfec* plants. **a**, **b**, Shannon indexes (**a**) and the relative abundance of bacteria at the phylum level (**b**), obtained from 16S rRNA gene-sequence profiles of endophytic leaf bacteria in plants grown in *Arabidopsis* mix soil supplemented with SynCom$^{Col\text{-}0}$. Data presentation and statistical analysis as in Fig. 1c, d. $n = 20$ (Col-0), $n = 19$ (*fec*), $n = 19$ (*min7*) and $n = 19$ (*mfec*) biological replicates passing quality control across 4 independent experiments. **c**, **d**, Population size of endophytic leaf microbiota (**c**) and leaf appearance (**d**) in 5-week-old plants 6 days after plants were shifted to high humidity (approximately 95%). One-way ANOVA with Tukey's test. Data are mean ± s.e.m., $n = 3$ biological replicates; experiments were repeated four times with similar results.

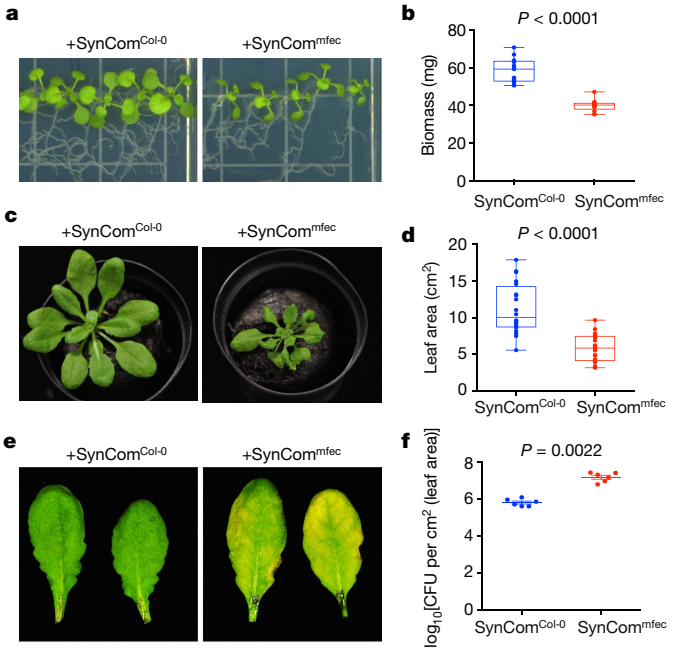


Fig. 3 | Functional effect of SynCom$^{Col\text{-}0}$ and SynCom^{mfec} on plant health. **a**, **b**, Phenotype (**a**) and biomass (**b**) of Col-0 seedlings inoculated with SynCom$^{Col\text{-}0}$ or SynCom^{mfec}. Twelve 14-day-old seedlings were weighed as one biological replicate (see Methods). $n = 13$ biological replicates. **c**, **d**, Appearance (**c**) and the total leaf area per plant (as one biological replicate) (**d**) of Col-0 plants grown in GnotoPots in the presence of SynCom$^{Col\text{-}0}$ or SynCom^{mfec} for 26 days. $n = 20$ biological replicates. **e**, **f**, Col-0 leaves were infiltrated with SynCom$^{Col\text{-}0}$ or SynCom^{mfec} by syringe injection of 1×10^8 CFU ml^{-1}, and leaf images (**e**) and bacterial populations (**f**) were recorded 5 days after infiltration. Data are mean ± s.e.m., $n = 6$ biological replicates. In box plots in **b** and **d**, the centre line is the median, box edges show the 75th and 25th percentiles, and whiskers cover the full range of values. In **b**, **d** and **f**, two-tailed Mann–Whitney *U*-test was used for statistical analysis. All experiments were repeated three times with similar results.

plant growth system (hereafter, GnotoPot; Methods) and found that the *mfec* plants appeared healthy in the absence of microbiota (Extended Data Fig. 2d, e). By contrast, we observed chlorosis and some necrosis in *mfec* plants in the presence of leaf endosphere-derived SynCom$^{Col\text{-}0}$, whereas wild-type Col-0 plants remained healthy in the presence of SynCom$^{Col\text{-}0}$. Thus, SynCom$^{Col\text{-}0}$ is sufficient to partially recapitulate host genotype-dependent dysbiotic symptoms in the phyllosphere.

Next, we addressed the question of whether the *mfec*-associated (that is, 'improperly assembled') microbiota alone is sufficient to cause dysbiotic symptoms in wild-type Col-0 plants. For this purpose, we assembled a 52-member *mfec* leaf-derived endophytic bacterial community (SynCom^{mfec}; Supplementary Table 1), which was prepared in parallel with SynCom$^{Col\text{-}0}$. Genome sequencing of individual isolates in SynCom$^{Col\text{-}0}$ and SynCom^{mfec} confirmed a more diverse and balanced bacterial composition in SynCom$^{Col\text{-}0}$ compared with SynCom^{mfec} (Extended Data Fig. 3a, Supplementary Tables 3–5), partially reflecting the endophytic bacterial composition found in Col-0 and *mfec* leaves as revealed by 16S rRNA gene sequencing (Fig. 1d). In particular, Firmicutes isolates were relatively abundant (20.8% of isolates) in SynCom$^{Col\text{-}0}$, whereas no culturable Firmicutes were recovered from SynCom^{mfec}. Conversely, 96.2% of isolates were Proteobacteria in SynCom^{mfec}, compared with 62.5% in SynCom$^{Col\text{-}0}$ (Supplementary Table 1). There were additional taxonomic differences in the two synthetic communities even though they were derived from Col-0 and *mfec* plants that were grown in the same soil and growth chamber at the same time (Methods), illustrating the powerful influence of the *mfec* genotype on the assembly of the leaf endophytic bacterial community.

We conducted three types of functional assays to rigorously test whether *mfec*-associated (that is, incorrectly assembled) microbiota could cause health-damaging dysbiosis. First, in Linsmaier–Skoog agar plate assays, Col-0 plants inoculated with SynCom^{mfec} had significantly reduced biomass relative to those inoculated with SynCom$^{Col\text{-}0}$ (Fig. 3a, b). Second, when grown in peat-based GnotoPots, Col-0 plants appeared healthy in the presence of SynCom$^{Col\text{-}0}$, but showed varying degrees of seedling stunting and an overall reduction in rosette size in the presence of SynCom^{mfec} (Fig. 3c, d). Third, when infiltrated at a concentration of 1×10^8 colony-forming units (CFU) per ml into otherwise healthy leaves of fully grown and colonized Col-0 plants, SynCom^{mfec}, but not SynCom$^{Col\text{-}0}$, induced prominent necrosis and chlorosis (Fig. 3e). In addition, SynCom^{mfec} grew to a higher population than SynCom$^{Col\text{-}0}$ in Col-0 leaves (Fig. 3f). When infiltrated at a lower concentration of 1×10^7 CFU ml^{-1} (equivalent to the approximately 10^4 CFU per mg (leaf tissue) of microbiota in *Arabidopsis*; Extended Data Fig. 3b), which simulates the level of endophytic microbiota in *mfec* leaves (Fig. 1b), SynCom^{mfec} was still capable of causing tissue damage, but to a lesser degree and more sporadically (Extended Data Fig. 3c). These combined results, from three independent assays, demonstrated that a dysbiotic microbiota (SynCom^{mfec}) is sufficient to confer a negative health effect in wild-type plants and provided evidence of the importance of assembling a normal leaf endophytic microbiota to ensure phyllosphere health.

Next, we investigated whether individual strains in SynCom$^{Col\text{-}0}$ and SynCom^{mfec} could cause tissue damage when infiltrated into leaves of Col-0 plants grown in *Arabidopsis* mix. With an inoculum containing 1×10^8 CFU ml^{-1}, more SynCom^{mfec} isolates (32 strains) than SynCom$^{Col\text{-}0}$ isolates (17 strains) caused tissue damage, supporting the hypothesis that *mfec* leaves are enriched for tissue-damaging bacteria (Supplementary Table 1). With an inoculum containing 1×10^7 CFU ml^{-1} (equivalent to

approximately 10^4 CFU per mg (leaf tissue), similar to the total endophytic microbiota in *mfec* leaves), ten SynCom^{mfec} strains, but only four SynCom$^{Col\text{-}0}$ strains, induced mild tissue damage (Supplementary Table 1). Of note, none of the 'symptom-inducer' strains multiplied at a rate similar to *Pst* DC3000, a virulent pathogen of *Arabidopsis* (Extended Data Fig. 4a), consistent with the hypothesis that these symptom-inducer strains are not canonical pathogens per se, but probably represent potentially harmful members of a normal leaf microbiota that are kept at low, non-damaging levels in a healthy wild-type phyllosphere. We further assembled a five-member synthetic phyllosphere community (SynCom$^{mix\ 5}$), consisting of Proteobacteria strains derived from SynCom^{mfec}, that induced robust leaf tissue damage (Extended Data Fig. 4b). Each of the five strains were sufficient to cause leaf damage on their own, suggesting functional redundancy within SynCom^{mfec} in the induction of phyllosphere dysbiosis. However, simple removal of Firmicutes from SynCom$^{Col\text{-}0}$ was not sufficient to produce a dysbiotic bacterial community (Extended Data Fig. 3c). This is consistent with the many fine taxonomical differences between strains in SynCom$^{Col\text{-}0}$ and SynCom^{mfec}.

Mechanisms of bacterial community shift

Next, we investigated the underlying mechanism by which *mfec* plants lost the ability to maintain endophytic leaf bacterial diversity. We hypothesized that, in addition to host genetic influences, antagonistic bacterial interactions might be involved. To test this hypothesis, we performed binary inhibition assays (2,116 combinations), on R2A medium, of 46 strains that represent all bacterial species we identified in SynCom$^{Col\text{-}0}$ and SynCom^{mfec}. This assay revealed a pattern of almost unidirectional antibiosis: most Firmicutes were strongly inhibited by a subset of Proteobacteria in vitro (Extended Data Fig. 5).

This in vitro observation was unexpected as it cannot explain the coexistence of Firmicutes and Proteobacteria in Col-0 leaves. We therefore considered the possibility that the largely unidirectional antibiosis observed on R2A medium (that is, in vitro) may become biologically relevant only when the two bacteria are in close proximity in vivo (that is, when endophytic bacterial populations become relatively high, as observed in *mfec* leaves; Fig. 1b). To investigate this possibility, we examined in vivo pairwise interactions of bacteria that showed strong binary interactions in vitro (for example, Proteobacteria strain C13 and Firmicutes strain C3, Proteobacteria strain C45 and Firmicutes strain C3, and Proteobacteria strain C13 and Firmicutes strain C41). These pairs were infiltrated into leaves of Col-0 plants grown in *Arabidopsis* mix potting soil at two different bacterial concentrations: 1×10^4 CFU ml^{-1} (equivalent to 1×10^2 CFU per cm^2 (leaf area)) and 1×10^6 CFU ml^{-1} (equivalent to 1×10^4 CFU cm^2 (leaf area)). For all three binary interactions examined, Firmicutes strains were outcompeted only at the higher inoculation level (Extended Data Fig. 6a, c, d). No such competition was observed between Proteobacterium strain C52 and Firmicutes strain C3, which did not show inhibition in vitro (Extended Data Figs. 5b, 6b). These results suggest a possible mechanism to explain why *mfec* plants lost the ability to maintain endophytic leaf bacterial diversity. In wild-type Col-0 leaves, pattern-triggered immunity and the MIN7 vesicle-trafficking pathway restrain the growth of leaf endophytic bacteria, including specific Proteobacteria that could inhibit other leaf endophytic bacteria, most notably, Firmicutes. In the *mfec* mutant, excess proliferation of leaf endophytic bacteria probably leads to the inhibition of Firmicutes by Proteobacteria strains, contributing to the reduction of overall relative bacterial diversity and conversion of a Firmicutes-rich community in Col-0 leaves to a Proteobacteria-rich community in the *mfec* mutant leaves.

A framework for microbiota homeostasis

Our results thus far provided strong evidence for the importance of pattern-triggered immunity and the MIN7 vesicle-trafficking pathway in controlling endophytic phyllosphere microbiota; however, it remained unclear whether these two processes are mechanistically separate or are components of a common molecular framework. Fortuitously, during this study we discovered *mfec*-like phenotypes in *ben3*, an *Arabidopsis* mutant that was initially isolated on the basis of a genetic screen for a defect in intracellular vesicle trafficking[20,21]. The *ben3* mutant carries a mutation in the *BREFELDIN A-INHIBITED GUANINE NUCLEOTIDE-EXCHANGE PROTEIN2* (*BIG2*) gene, which encodes an ADP ribosylation factor (ARF) family of guanine nucleotide exchange factor that is closely related to MIN7 and, like MIN7, is localized in the *trans*-Golgi network and early endosome[21]. The *ben3* mutant phenocopied the *mfec* quadruple mutant in (1) exhibiting spontaneous dysbiosis-like symptoms (Fig. 4a, Extended Data Fig. 7a, c), (2) harbouring a higher level of leaf endophytic microbiota compared with Col-0 plants (Fig. 4b, Extended Data Fig. 7b) and (3) hosting a leaf endophytic microbial community that is reduced in Shannon diversity index, enriched in Proteobacteria and depleted in Firmicutes (Fig. 4c, Extended Data Fig. 9a). However, during further characterization, we found that independent *big2* mutants carrying transfer DNA (T-DNA) insertions (*big2-1* and *big2-2*) did not show dysbiosis phenotypes (Extended Data Fig. 8a, b, g) and, using bulk segregation analysis and next-generation sequencing (Methods), we identified the causal mutation of dysbiosis in *ben3* to be a S205F substitution in a MACPF-domain protein—encoded by the *CONSTITUTIVELY ACTIVATED CELL DEATH1* (*CAD1*) gene (Extended Data Fig. 8c–g, Supplementary Table 6)—localized to the plasma membrane[22–24].

Because the *cad1*^{S205F} mutation phenocopied the *mfec* quadruple mutants that are defective in both pattern-triggered immunity and the MIN7 vesicle-trafficking pathway, we hypothesized that CAD1 could be one of the convergent components downstream of pattern-triggered immunity and the MIN7 vesicle-trafficking pathway. Consistent with

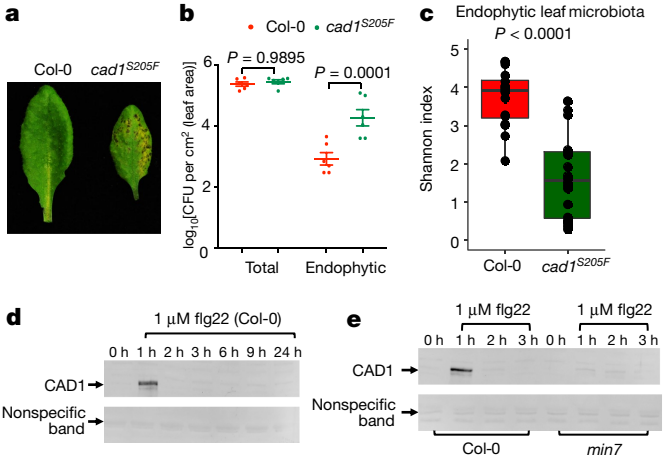


Fig. 4 | Microbiota phenotypes in the *ben3* mutant. a, **b**, Leaf appearance (**a**) and population sizes of total and endophytic leaf microbiota (**b**) in Col-0 and *ben3* (hereafter referred to as *cad1*^{S205F}) plants grown in *Arabidopsis* mix soil supplemented with SynCom$^{Col\text{-}0}$ for 4 weeks before plants were shifted to high humidity (approximately 95%) for 2 days (see Methods). One-way ANOVA with Tukey's test. Data are mean ± s.e.m., $n = 6$ biological replicates. **c**, Shannon indexes of 16S rRNA gene-sequence profiles of endophytic leaf bacteria in Col-0 and *cad1*^{S205F} plants supplemented with SynCom$^{Col\text{-}0}$. Data presentation and statistical analysis as in Fig. 1c. $n = 20$ (Col-0) and $n = 20$ (*cad1*^{S205F}) biological replicates. **d**, **e**, Western blot analyses of CAD1 protein in Col-0 (**e**) and *min7* (**f**) plants. Five-week-old Col-0 and *min7* leaves were infiltrated with 1 μM flg22 and collected at the indicated time points. CAD1 protein was detected with a CAD1 antibody; nonspecific bands show equal loading. The uncropped gel images are shown in Supplementary Fig. 1. All experiments in this figure were repeated three times with similar results.

this possibility, we found that the *CAD1* gene and protein were induced in response to flg22 (Fig. 4d, Extended Data Fig. 8h), an inducer of pattern-triggered immunity[25,26], and that flg22-induced accumulation of the CAD1 protein was affected by the *min7* mutation (Fig. 4e, Extended Data Fig. 8i). These results suggest that pattern-triggered immunity, the MIN7 vesicle-trafficking pathway and CAD1 are components of a large molecular framework that controls endophytic microbial abundance and diversity in the phyllosphere (Extended Data Fig. 9b, c).

Discussion

Our results begin to highlight conceptual parallels between plants and mammals in the mechanisms that prevent dysbiosis, a condition with severe consequences for the health of the host. In particular, defects in innate immunity pathways seem to be a common determinant underlying dysbiosis in mammals[27–29] and plants. In addition, CAD1 is a member of the MACPF protein family; members of this family, such as complement protein C9 and perforin, have been shown to be involved in innate and adaptive immunity against pathogens in mammals[30,31]. Pattern-triggered immunity, MIN7 and CAD1 are broadly conserved across plant taxa (Extended Data Fig. 10, Supplementary Table 7), suggesting that host control of the endophytic phyllosphere population and diversity may be a conserved feature in the plant kingdom. Manipulation of host genetic pathways regulating microbiota homeostasis could lead to a more beneficial and climate-resilient phyllosphere microbiota, which could in turn improve the performance of natural ecosystems and agricultural crops, as discussed further in the Supplementary Discussion.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2185-0.

1. Xin, X. F. et al. Bacteria establish an aqueous living space in plants crucial for virulence. *Nature* **539**, 524–529 (2016).
2. Beattie, G. A. & Lindow, S. E. Bacterial colonization of leaves: a spectrum of strategies. *Phytopathology* **89**, 353–359 (1999).
3. Lindow, S. E. & Brandl, M. T. Microbiology of the phyllosphere. *Appl. Environ. Microbiol.* **69**, 1875–1883 (2003).
4. Berendsen, R. L., Pieterse, C. M. & Bakker, P. A. The rhizosphere microbiome and plant health. *Trends Plant Sci.* **17**, 478–486 (2012).
5. Xu, L. et al. Drought delays development of the sorghum root microbiome and enriches for monoderm bacteria. *Proc. Natl Acad. Sci. USA* **115**, E4284–E4293 (2018).
6. Edwards, J. A. et al. Compositional shifts in root-associated bacterial and archaeal microbiota track the plant life cycle in field-grown rice. *PLoS Biol.* **16**, e2003862 (2018).
7. Finkel, O. M., Castrillo, G., Herrera Paredes, S., Salas González, I. & Dangl, J. L. Understanding and exploiting plant beneficial microbes. *Curr. Opin. Plant Biol.* **38**, 155–163 (2017).
8. Pieterse, C. M. J., de Jonge, R. & Berendsen, R. L. The soil-borne supremacy. *Trends Plant Sci.* **21**, 171–173 (2016).
9. Duran, P. et al. Microbial interkingdom interactions in roots promote *Arabidopsis* survival. *Cell* **175**, 973–983 (2018).
10. Müller, D. B., Vogel, C., Bai, Y. & Vorholt, J. A. The plant microbiota: systems-level insights and perspectives. *Annu. Rev. Genet.* **50**, 211–234 (2016).
11. Zhang, J. et al. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. *Nat. Biotechnol.* **37**, 676–684 (2019).
12. Horton, M. W. et al. Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. *Nat. Commun.* **5**, 5320 (2014).
13. Bodenhausen, N., Bortfeld-Miller, M., Ackermann, M. & Vorholt, J. A. A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. *PLoS Genet.* **10**, e1004283 (2014).
14. Wagner, M. R. et al. Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nat. Commun.* **7**, 12151 (2016).
15. Laforest-Lapointe, I., Paquette, A., Messier, C. & Kembel, S. W. Leaf bacterial diversity mediates plant diversity and ecosystem function relationships. *Nature* **546**, 145–147 (2017).
16. Tang, D., Wang, G. & Zhou, J. M. Receptor kinases in plant–pathogen interactions: more than pattern recognition. *Plant Cell* **29**, 618–637 (2017).
17. Boutrot, F. & Zipfel, C. Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annu. Rev. Phytopathol.* **55**, 257–286 (2017).
18. Turpin, W., Goethel, A., Bedrani, L. & Croitoru, K. Determinants of IBD heritability: genes, bugs, and more. *Inflamm. Bowel Dis.* **24**, 1133–1148 (2018).
19. Sokol, H. & Seksik, P. The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. *Curr. Opin. Gastroenterol.* **26**, 327–331 (2010).
20. Tanaka, H., Kitakura, S., De Rycke, R., De Groodt, R. & Friml, J. Fluorescence imaging-based screen identifies ARF GEF component of early endosomal trafficking. *Curr. Biol.* **19**, 391–397 (2009).
21. Kitakura, S. et al. BEN3/BIG2 ARF GEF is involved in brefeldin A-sensitive trafficking at the *trans*-Golgi network/early endosome in *Arabidopsis thaliana*. *Plant Cell Physiol.* **58**, 1801–1811 (2017).
22. Morita-Yamamuro, C. et al. The *Arabidopsis* gene *CAD1* controls programmed cell death in the plant immune system and encodes a protein containing a MACPF domain. *Plant Cell Physiol.* **46**, 902–912 (2005).
23. de Michele, R. et al. Free-flow electrophoresis of plasma membrane vesicles enriched by two-phase partitioning enhances the quality of the proteome from *Arabidopsis* seedlings. *J. Proteome Res.* **15**, 900–913 (2016).
24. Alexandersson, E., Saalbach, G., Larsson, C. & Kjellbom, P. *Arabidopsis* plasma membrane proteomics identifies components of transport, signal transduction and membrane trafficking. *Plant Cell Physiol.* **45**, 1543–1556 (2004).
25. Felix, G., Duran, J. D., Volko, S. & Boller, T. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *Plant J. Cell Mol. Biol.* **18**, 265–276 (1999).
26. Zipfel, C. et al. Bacterial disease resistance in *Arabidopsis* through flagellin perception. *Nature* **428**, 764–767 (2004).
27. Levy, M., Kolodziejczyk, A. A., Thaiss, C. A. & Elinav, E. Dysbiosis and the immune system. *Nat. Rev. Immunol.* **17**, 219–232 (2017).
28. Hall, A. B., Tolonen, A. C. & Xavier, R. J. Human genetic variation and the gut microbiome in disease. *Nat. Rev. Genet.* **18**, 690–699 (2017).
29. Sun, L., Nava, G. M. & Stappenbeck, T. S. Host genetic susceptibility, dysbiosis, and viral triggers in inflammatory bowel disease. *Curr. Opin. Gastroenterol.* **27**, 321–327 (2011).
30. McCormack, R., de Armas, L., Shiratsuchi, M. & Podack, E. R. Killing machines: three pore-forming proteins of the immune system. *Immunol. Res.* **57**, 268–278 (2013).
31. Spicer, B. A., Conroy, P. J., Law, R. H. P., Voskoboinik, I. & Whisstock, J. C. Perforin—a key (shaped) weapon in the immunological arsenal. *Semin. Cell Dev. Biol.* **72**, 117–123 (2017).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Plant materials and growth conditions

For most experiments, *Arabidopsis thaliana* plants were grown in *Arabidopsis* mix greenhouse potting soil (equal parts of Suremix (Michigan Grower Products), medium vermiculate and perlite; autoclaved once) or Redi-Earth soil (Sun Gro Horticulture) in air-circulating growth chambers for colonization of phyllosphere microbiota. Plants were grown under relative humidity set at 60%, temperature at 22 °C, light intensity at 100 μE m^{-2} s^{-1} and photoperiod at a 12:12 h light:dark cycle. Five-week-old plants were used for bacterial inoculation and microbiota assays.

The wild-type accession Col-0, *min7 fls2 efr cerk1* (*mfec*), *ben3* (that is, *cad1*^{S205F} used in this study) and *dde2-2/ein2-1/pad4-1/sid2-2* (*deps*) mutant derivatives were described previously[1,21,32] or in this study. The *big2-1* (SALK_033446) and *big2-2* (SALK_016558) T-DNA insertion mutants were obtained from the *Arabidopsis* Biological Resource Centre (ABRC) at The Ohio State University.

Bacterial quantification

To quantify culturable endophytic bacterial communities, leaves were surface-sterilized in 75% ethanol for 1 min and rinsed in sterile water twice. After air-drying to evaporate surface water, leaves were weighed and ground in 1 ml sterile 10 mM $MgCl_2$ buffer. A serial dilution in 100 μl (to up to 10^{-3} for Col-0 and up to 10^{-5} for *mfec* and *cad1*^{S205F} mutants) was made and plated on R2A plates (100 mm × 100 mm). Under experimental conditions in this study, almost all leaf bacteria grew to large enough microcolonies in R2A plates in 2–3 days to be efficiently counted under a microscope (20× magnification). Pilot trials with longer incubation times, up to 10 days (see below), did not yield significantly more bacterial colonies that would have affected plant genotype differences. Therefore, we counted bacterial colony-forming units (CFU) 3 days after R2A plates were incubated at room temperature. CFUs were normalized to fresh tissue weight or leaf disk area. Total leaf bacterial communities were determined following the same protocol except without surface sterilization. A serial dilution was made in 100 μl volume (up to 10^{-5} for Col-0, *mfec* and *cad1*^{S205F}) and plated on R2A plates.

In pilot trials, two leaves or eight leaf disks were ground in 1 ml sterile 10 mM $MgCl_2$ buffer. After serial dilutions (see above), 100 μl of leaf homogenate were spread onto square Petri dishes (100 mm × 100 mm) containing 30 ml R2A agar. Degrees of dilution were determined by empirical experiments to ensure that there were 20 to 200 CFUs per plate for accurate counting. As an example, when the following two incubation methods were used: (i) 3 days of incubation at room temperature and (ii) 10 days of incubation (2 days at room temperature, 6 days at 4 °C and 2 days at room temperature), bacterial CFUs (mean ± s.e.m., $n = 4$) were 105 ± 20 (2-day incubation) and 101 ± 23 (10-day incubation), respectively.

Gnotobiotic plant growth systems

Three gnotobiotic plant growth systems were used in this study. Murashige–Skoog (MS) or Linsmaier–Skoog (LS) agar plate system: Col-0, *mfec* and *cad1*^{S205F} seeds were surface-sterilized, cold-stratified and germinated on 0.5× MS or LS agar medium plates. Seedlings were grown under 50 μE $m^{-2}s^{-1}$ with a photoperiod of 12:12 h light:dark cycle. FlowPot gnotobiotic system: a peat-based gnotobiotic system described previously[33]. GnotoPot: a compressed peat pellet-based gnotobiotic system developed in this study as a simpler, alternative potting soil-based gnotobiotic system. In brief, compressed peat pellets (Jiffy Products) were transferred to 2-inch polypropylene pots and hydrated to saturation with LS medium buffered with 2-(*N*-morpholino) ethanesulfonic acid (MES) to pH 5.7 (Caisson Labs). GnotoPots were then placed in plant tissue culture microbox (SacO2) that had a no. 40 green filter mounted in the lid. Assembled microboxes with lids loosely placed were placed inside an autoclave bag (Sun Bag, Sigma) and autoclaved twice for 45 min each, with 24 h storage at room temperature in between. After GnotoPots were cooled down, microboxes were sealed and stored until time of use. Surface-sterilized *Arabidopsis* seeds were stratified at 4 °C for 24 h before being sown into GnotoPots under germ-free conditions. GnotoPots were then placed in a tissue culture growth chamber set at 22 °C, 12:12 h light:dark cycle photoperiod and light intensity of 100 μE m^{-2} s^{-1}. Sterility of FlowPots and GnotoPots was routinely monitored by plating samples of plants and peat substrate in R2A plates.

For induction of dysbiosis symptoms shown in Extended Data Fig. 2e, 5.5- to 6.5-week-old GnotoPots-grown plants were sprayed with sterile water, placed under a clear plastic dome to achieve high humidity (~95%) under 40 μE m^{-2} s^{-1} light intensity and temperature at 23 °C for 10 days. For results shown in Fig. 3c, plants were grown in GnotoPots under a photoperiod of 16:8 h light:dark (100 μE m^{-2} s^{-1} and temperature at 22 °C) for 26 days and total leaf areas were measured with the Easy Leaf Area software[34].

Synthetic communities of leaf endophytic bacteria

To generate SynComs, Col-0 and *mfec* plants were grown in *Arabidopsis* mix potting soil to 5 weeks old and were sprayed with water and kept under high humidity (~95%) for 5 days. Representative leaves were harvested (8 leaves were picked from 4 plants of each genotype) and surface-sterilized in 75% ethanol for 1 min and rinsed in sterile water twice. Leaves were ground in sterile water, and bacterial suspensions were diluted (to 10^{-3} for Col-0 and 10^{-5} for *mfec*) and plated on R2A plates, which were kept at 22 °C for 4 days. About 50 colonies from each genotype were randomly picked, constituting SynCom$^{Col\text{-}0}$ and SynCom^{mfec}, respectively.

For addition of SynCom bacteria as internal control to *Arabidopsis* potting soil for 16S rRNA gene-sequence profiles (Figs. 2, 4 and Extended Data Figs. 2, 9), individual bacterial strains were scraped from R2A plate and suspended in 10 mM $MgCl_2$ buffer, bacterial suspensions were adjusted to the same OD_{600}. Equal volumes of each strain were pooled and diluted to a final OD_{600} of 0.01 (~0.5×10^7 CFU ml^{-1}). Five millilitres of prepared SynCom was added to each *Arabidopsis* mix soil pot. To reduce background microbiota that is naturally present in *Arabidopsis* mix soil, *Arabidopsis* mix soil and meshes were autoclaved twice before addition of SynCom bacteria; pots, flats and plastic domes used in growing plants were sprayed with 75% ethanol. Surface-sterilized Col-0, *mfec*, *min7*, *fec* or *cad1*^{S205F} seeds were then added to the assembled soil pots. Plants were watered with autoclaved nutrient water 1–2 times each week.

For experiments with 0.5× LS agar plates (Fig. 3a, b), sterile Col-0 seeds were germinated on 0.5× LS plates (without sucrose) in the presence of 2 μl 10^7 CFU ml^{-1} SynCom$^{Col\text{-}0}$ or SynCom^{mfec} for 14 days. The 2 μl SynCom was applied directly to each seed. For preparation of SynCom$^{Col\text{-}0}$ and SynCom^{mfec} for inoculation into FlowPots or GnotoPots (Fig. 3c, d and Extended Data Figs. 2, 7), SynCom$^{Col\text{-}0}$ and SynCom^{mfec} mixtures were prepared as above and the final OD_{600} was adjusted to 0.04 (~2×10^7 CFU ml^{-1}). A single seed was sown to each pot and 1 ml of the SynCom suspension was added. For experiments to compare the effects of different SynComs on leaf health grown in *Arabidopsis* potting soil (Fig. 3e, f and Extended Data Fig. 3b, c), SynCom$^{Col\text{-}0}$ and SynCom^{mfec} mixtures were prepared as above, and the final OD_{600} was adjusted to 0.2 (~1×10^8 CFU ml^{-1} for total SynCom mixtures; ~2×10^6 CFU ml^{-1} for each strain, Fig. 3e, f) or 0.02 (~1×10^7 CFU ml^{-1} for total SynCom mixtures; ~2×10^5 CFU ml^{-1} for each strain, Extended Data Fig. 3b, c) before infiltration into 4-week-old leaves.

16S rRNA gene-sequence profiling

Col-0, *mfec*, *min7*, *fec* or *cad1*^{S205F} plants were grown on experimental *Arabidopsis* mix soil with or without inoculated SynCom$^{Col\text{-}0}$ added as described above. Five-week-old healthy plants were sprayed with water and kept under high humidity (~95%) for several days until dysbiosis symptoms appeared. Middle-age leaves from Col-0 and mutants were collected. To analyse endophytic bacterial community, leaves were first surface-sterilized with 5% (v/v) bleach for 1 min, followed with rinse with sterile ddH_2O twice. After blot-drying to remove surface water, two leaves from the same plant were collected in one tube as one biological replicate, which was then snap-frozen in liquid N_2 and stored at −80 °C. For analysis of total bacterial community, two leaves from one plant was collected in a tube as one sample, which was then snap-frozen in liquid N_2 and stored at −80 °C.

To prepare DNA for bacterial 16S rRNA gene-based community analysis, Total DNA from leaf samples was extracted using MoBio Power Soil DNA Isolation kit (Qiagen). PCR was performed using AccuPrime high-fidelity Taq DNA polymerase (Invitrogen) using barcoded primers 799F/1193R[35]. 799F: **ACACTGACGACATGGTTCTACA**AACMGGATTAGATACCCKG and 1193R: **TACGGTAGCAGAGACTTGGTCT**ACGTCATCCCCACCTTCC (bold sequences are the Illumina common sequence adapters). PCR was performed in triplicate in 25 µl reaction volumes containing 0.15 µl AccuPrime high-fidelity Taq DNA polymerase, 1 µl DMSO, 2.5 µl Buffer II, 0.5 µl of each primer (10 µM), 2 µl template DNA and 18.35 µl ddH_2O. The PCR program included a hot start at 94 °C for 60 s, 35 cycles of denaturation at 94 °C for 20 s, primer annealing at 53 °C for 30 s and extension at 68 °C for 45 s, followed by a final extension at 68 °C for 2 min and a cool down to 8 °C. PCR products were separated on 1% agarose gel. The 450-bp band of amplified bacterial 16S rRNA gene was extracted using Zymoclean Gel DNA Recovery Kit (Zymo Research) according to the manufacturer's instructions. DNA concentration was measured with PicoGreen dsDNA assay kit (Life Technologies) and adjusted to 1 to 10 ng µl^{-1} for all samples. Samples were submitted to Research Technology Service Facility (RTSF) at Michigan State University for library preparation and 16S rRNA gene sequencing (see below).

RTSF Genomics Core at Michigan State University completed library preparation by PCR with dual-indexed Illumina-compatible adapters targeting the Fluidigm barcoding oligos. Final PCR products were bulk-normalized using Invitrogen SequalPrep DNA Normalization plates and recovered libraries were pooled. The library pool was cleaned up using AmpureXP magnetic beads and then quantified using a combination of Qubit dsDNA HS (Invitrogen). Agilent 4200 TapeStation DNA 1000 and Kapa Illumina Library Quantification qPCR assays: the library pool was loaded onto an Illumina MiSeq Standard v2 flow cell and sequencing was performed in a 2 × 250 bp paired-end format using a MiSeq v2 500 cycle reagent cartridge. Common sequencing and index primers complementary to the Fluidigm CS1/CS2 oligomers were added to appropriate wells of the reagent cartridge. Base calling was done by Illumina Real Time Analysis (RTA) v.1.18.54 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2fastq v.2.19.1.

Raw Illumina fastq files were quality-filtered and taxonomically analysed using QIIME 2 Core 2018.11 distribution[36]. In brief, primers of imported sequences were removed via Cutadapt[37]. DADA2[38] was used to filter and denoise sequences, remove chimaeras, identify representative sequences of OTUs and create an OTU table. Representative sequences of OTUs were taxonomically annotated using a pre-trained naive Bayes classifier[39] on the basis of the bacterial 16S rRNA Greengenes reference database (13_8 release). From this taxonomic annotation, all unassigned sequences and sequences annotated as mitochondria and chloroplast were removed. The filtered sequences were clustered at 97% similarity and the resulting OTU table was then used to determine taxonomic distributions and alpha (observed OTUs and Shannon's species diversity index). For alpha-diversity calculations, samples were rarefied to the same number of reads.

The total bacterial community in Col-0 leaves at phylum level are dominated by Proteobacteria (92.2% relative abundance), Actinobacteria (5%), Firmicutes (1.1%) and Bacteroidetes (0.7%) in this study, which are similar (Pearson's correlation = 0.96, P = 0.037) to the *Arabidopsis* leaf natural communities found by Bai and colleagues[35]. At the order level, we found that 9 of the top-10 orders (Burkholderiales, Actinomycetales, Methylophilales, Sphingomonadales, Rhizobiales, Pseudomonadales, Enterobacteriales, Flavobacteriales and Caulobacterales) in our total Col-0 leaf bacterial community are also identified as the top-10 orders of *Arabidopsis* leaf natural communities found by Bai and colleagues[35].

Bacterial genome assembly and taxonomic classification

To isolate high molecular weight genomic DNA of bacterial isolates in SynCom$^{Col\text{-}0}$ and SynCom^{mfec}, we used a modified EZNA Bacterial DNA Kit Protocol (OMEGA Bio-Tek, UAS). Quality of genomic DNA was analysed by agarose gel (1% (w/v)) and quantified by Qubit. Approximately 1 µg of genomic DNA was used for Oxford Nanopore bacterial sequencing and 30 µg of genomic DNA for Illumina sequencing.

Using the software Canu v.1.7 with default parameters[40], raw nanopore reads were corrected, trimmed and then assembled into long contigs. A majority of the assembled genomes of SynCom$^{Col\text{-}0}$ and SynCom^{mfec} are complete or near complete, with genome size ranging from 3.54 to 9.42 Mb (Supplementary Table 3). To further improve the quality of the assembled genomes, we resequenced full genomes for 27 relevant strains using the Illumina sequencing platform. After removing adaptor sequences, trimming and removing low-quality reads using the software Sickle[36] v.1.33, more than 24 Gb high-quality short reads were generated (Supplementary Table 3). We applied high-quality short reads from Illumina sequencing platform to correct the assembled genomes using program Pilon[41] v.1.22. To generate accurate taxonomic information for the sequenced genomes, the average nucleotide identity of the whole genome with references and phylogenetic relationship with references on the basis of 120 marker genes were inferred using program gtdbtk[42] v.0.1.3. We also used Ribosomal Database Project tools to infer the taxonomy of sequenced isolates based on full-length 16S rRNA genes using software Mothur[43] v.1.34.2. The maximum-likelihood phylogenetic tree for sequenced isolates were constructed based on the full-length 16S rRNA gene using MEGA7[44]. A total of 100 bootstrap replicates were made.

Binary interaction

Bacterial strains were individually cultured on R2A plates at 28 °C for 1–2 days. One full inoculation loop of bacteria was suspended in 3 ml R2A medium (OD_{600} of 2.1 to 4.0). For making a bacterial lawn, 2.6 ml of a 'target' bacterial suspension was added to 40 ml of molten R2A agar pre-cooled to 42 °C, gently mixed and then poured into two square Petri dishes (100 × 100 mm). Two microlitres of each 'attacker' strain was spotted onto the plate and incubated at room temperature for 3 days at which photographs were taken to observe inhibition zones. Two technical repeats were performed for each strain, strain had two technical repeats, and experiment was conducted three times. In total, 2,116 binary interactions were examined.

Characterization of the *ben3* mutant

The *ben3* (*cad1*^{S205F}) mutant was backcrossed to wild type Col-0 to generate a mapping-by-sequencing population. Of 24 F1 plants grown in soil, all showed wild type Col-0 phenotype (that is, no spontaneous dysbiosis symptoms), indicating mutant phenotype is a recessive trait. F1 plants were allowed to self and produce a segregating F2 population. Of 376 F2 plants, 88 showed mutant phenotype (that is, spontaneous dysbiosis symptoms), whereas the rest (288 plants) were similar to the wild type. The observed ~1:3 (mutant:wild type) phenotype segregation

ratio suggests the mutant phenotype is caused by a single mutation on a nuclear gene. To identify the causative mutation, genomic DNA was extracted from 50 F2 plants exhibiting mutant phenotype and 50 F2 plants exhibiting wild-type phenotype and pooled into mutant-like and wild type-like pools. Pooled DNA samples were submitted to the Michigan State University RTSF Genomic Core facility for library preparation (Illumina TruSeq Nano DNA Library Preparation Kit) and sequenced on Illumina HiSeq 4000 platform in a 2 × 150 bp paired-end-sequencing format. 70.3 and 94.3 million reads were obtained for mutant-like and wild-type-like pools, respectively. Whole-genome resequencing data were analysed following methods developed by Austin and colleagues[45], with minor package and version changes. Adaptor sequences and poor-quality reads were trimmed off using Trimmomatic (v.0.33). Reads were aligned to *Arabidopsis* TAIR10 genome using bowtie2 (v.2.3.1). Alignments were coordinate-sorted using SAMtools (v.1.5); PCR duplicate reads were removed using picardTools (v.1.89). Variances were called using bcftools (v.1.2). Candidate causative mutations were analysed using SHOREmap v.3.0[46] with algorithms developed for recessive mutation within a backcrossing population.

Production of *35S::CAD1* transgenic *Arabidopsis*

The *CAD1* coding sequence was amplified by PCR using the following primers. Sense primer: CACC**ATG**GAGAATCGTAAAGGAGGAAACT (start codon in bold); antisense primer: **TCA**ATAATTTAGCAACGAATACTTC (stop codon in bold). The amplified *CAD1* fragment was cloned into pENTR/D cloning vector (Invitrogen), and transferred by LR recombination into the binary expression vector pB7-35S::His-Flag-GW[47] to generate a *35S::His-Flag-CAD1* (named *35S::CAD1* for short). The binary vectors containing *35S::CAD1* gene was introduced into *Agrobacterium tumefaciens* C58C1 by electroporation. *Arabidopsis* plants were transformed using the floral-dip method[48]. Glufosinate ammonium (Basta) was used to select for transgenic T1 plants, which were further screened by western blot using a CAD1-specific antibody. Homozygous T3 plants expressing fusion proteins were used for analyses. CAD1-specific antibody was prepared against the C-terminal 240 amino acids. The NdeI–XhoI fragment of $CAD1_{322-561}$ was cloned into the pET28a vector (Novagen) to overexpress a His_6–CAD1 fusion protein. Sense primer: **CATATG**TGGGCTCCCGAACAGAGTAACCTCC (NdeI site bold, start codon underlined); antisense primer: **CTCGAG**TCAATAATTTAGCAACG AATACTTC (XhoI site bold, stop codon underlined). Guinea pigs were injected with purified His_6–CAD1 protein to raise CAD1-specific antibody (Cocalico Biologicals).

Data analysis, statistics and experimental repeats

Plants of different genotypes (Col-0, *mfec* and *cad1*^{S205F}) were grown side by side to minimize unexpected environmental variations during growth and experimentation. Leaf samples of similar ages were collected and assessed randomly for each genotype. Researchers were not blinded to allocation during experiments and outcome assessment. This is in part because different plant genotypes under study (Col-0, *mfec* and *cad1*^{S205F}) exhibit very visually distinct phenotypes, making blinding not possible. Routine practices included more than one author observing and assessing phenotypes whenever possible. The specific statistical method used, the sample size, the number of experimental repeats and the results of statistical analyses are described in the relevant figure legends. Sample size was determined on the basis of experimental trials and with consideration of previous publications on similar experiments to allow for confident statistical analyses. Two-tailed Mann–Whitney *U*-test or one-way or two-way ANOVA with Tukey's test was used for multiple comparisons within a dataset, with significance at $P < 0.05$. ANOVA was performed with GraphPad Prism software. Statistical significance of alpha-diversity between plant genotypes were determined via Mann–Whitney *U*-test. The Benjamini–Hochberg method[49] was applied to correct the *P* values after performing multiple comparisons. Differential ASVs representing unique bacterial 16S rRNA sequences of endophytic phyllosphere microbiota between Col-0 and *mfec* mutants grown in potting soil supplemented with SynCom strains were identified with a negative binomial generalized linear model (GLM) in the edgeR package[50]. The Benjamini–Hochberg method false discovery rate was applied to correct the *P* values after performing multiple comparisons. ASVs with false discovery rate below or equal to 0.05 were considered differentially colonized (that is, enriched or depleted in *mfec* compared to Col-0).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw source 16S rRNA gene sequences from this project are available in the Sequence Read Archive database under BioProject PRJNA554246, accession numbers SAMN12259846 to SAMN12260169. Bacterial genome source data are available in the Sequence Read Archive database under the BioProject PRJNA555902. Source Data for Figs. 1–4 and Extended Data Figs. 3, 4, 6–8 are provided with the paper.

Code availability

Scripts used in the microbiota analyses are available at https://github.com/godlovexiaolin/A-genetic-network-for-host-control-of-phyllosphere-microbiota-for-plant-health. All other software used in this study are cited in the text.

32. Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J. & Katagiri, F. Network properties of robust immunity in plants. *PLoS Genet*. **5**, e1000772 (2009).
33. Kremer, J. M. P. et al. FlowPot axenic plant growth system for microbiota research. Preprint at *bioRxiv* https://doi.org/10.1101/254953 (2018).
34. Easlon, H. M. & Bloom, A. J. Easy Leaf Area: automated digital image analysis for rapid and accurate measurement of leaf area. *Appl. Plant Sci*. **2**, 1400033 (2014).
35. Bai, Y. et al. Functional overlap of the *Arabidopsis* leaf and root microbiota. *Nature* **528**, 364–369 (2015).
36. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol*. **37**, 852–857 (2019).
37. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. **17**, ej.17.1.200 (2011).
38. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
39. Werner, J. J. et al. Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys. *ISME J*. **6**, 94–103 (2012).
40. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res*. **27**, 722–736 (2017).
41. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
42. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol*. **36**, 996–1004 (2018).
43. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol*. **75**, 7537–7541 (2009).
44. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol*. **33**, 1870–1874 (2016).
45. Austin, R. S., Chatfield, S. P., Desveaux, D. & Guttman, D. S. Next-generation mapping of genetic mutations using bulk population sequencing. *Methods Mol. Biol*. **1062**, 301–315 (2014).
46. Sun, H. & Schneeberger, K. SHOREmap v3.0: fast and accurate identification of causal mutations from forward genetic screens. *Methods Mol. Biol*. **1284**, 381–395 (2015).
47. Lee, C. M., Adamchek, C., Feke, A., Nusinow, D. A. & Gendron, J. M. Mapping protein–protein interactions using affinity purification and mass spectrometry. *Methods Mol. Biol*. **1610**, 231–249 (2017).
48. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J. Mol. Cell Biol*. **16**, 735–743 (1998).
49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
50. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
51. Nomura, K. et al. A bacterial virulence protein suppresses host innate immunity to cause plant disease. *Science* **313**, 220–223 (2006).
52. Nomura, K. et al. Effector-triggered immunity blocks pathogen degradation of an immunity-associated vesicle traffic regulator in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **108**, 10774–10779 (2011).

Acknowledgements We thank members, including B. Hsueh, of the He laboratory for technical help and insightful discussions; T. Gong for help with testing leaf necrosis phenotypes of plants grown on MS agar plates; and H. Tanaka and J. Friml for *ben3/cad1*^{S205F}seeds. This project was supported by funding from National Institutes of Health (GM109928), the Department of Energy (the Chemical Sciences, Geosciences, and Biosciences Division, Office of Basic Energy Sciences, Office of Science; DE–FG02–91ER20021 for *ben3/cad1*^{S205F} mutant characterization) and Plant Resilience Institute at Michigan State University for support of optimization of the GnotoPot system (to S.Y.H.) and by funding from the CAS Center for Excellence in Molecular Plant Sciences and the Institute of Plant Physiology and Ecology, Chinese Academy of Sciences (to X.-F.X.).

Author contributions X.-F.X. and S.Y.H. conceptualized, designed the experiments and co-supervised the project. T.C. and K.N. performed most of experiments; X.-F.X. performed initial 16S sequencing set up and sample collection while at Michigan State University; R.S. performed GnotoPot experiments; X.W. performed 16S bioinformatics analysis; J.X. performed bacterial genome analysis; L.Y. performed the MS plate assay for Col-0 and the *mfec* mutant; B.C.P. performed 16S bioinformatics analysis. L.M. was involved in *cad1*-related experiments; J.K. was involved in initial 16S RNA gene sequencing design; Y.C. was involved in mapping the *cad1* mutation; L.Z. performed phylogenetic analysis of *CAD1* and *MIN7* genes and advised on statistical analyses; N.W. and E.W. advised on bioinformatics and statistical analyses. T.C., X.-F.X. and S.Y.H. wrote the manuscript with input from all co-authors. X.W. and R.S. contributed equally as co-second authors.

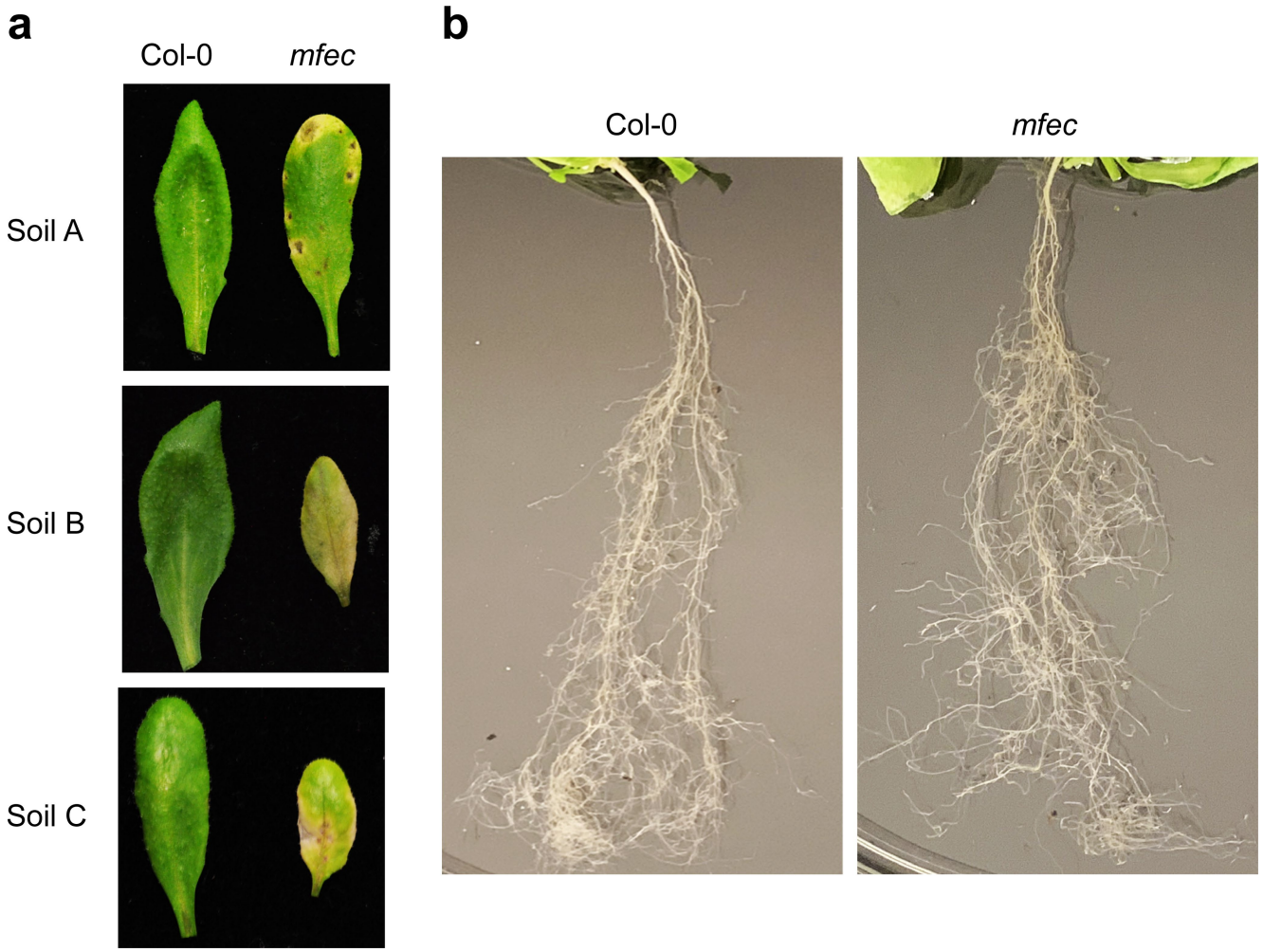
Competing interests The authors declare no competing interests.

Additional information

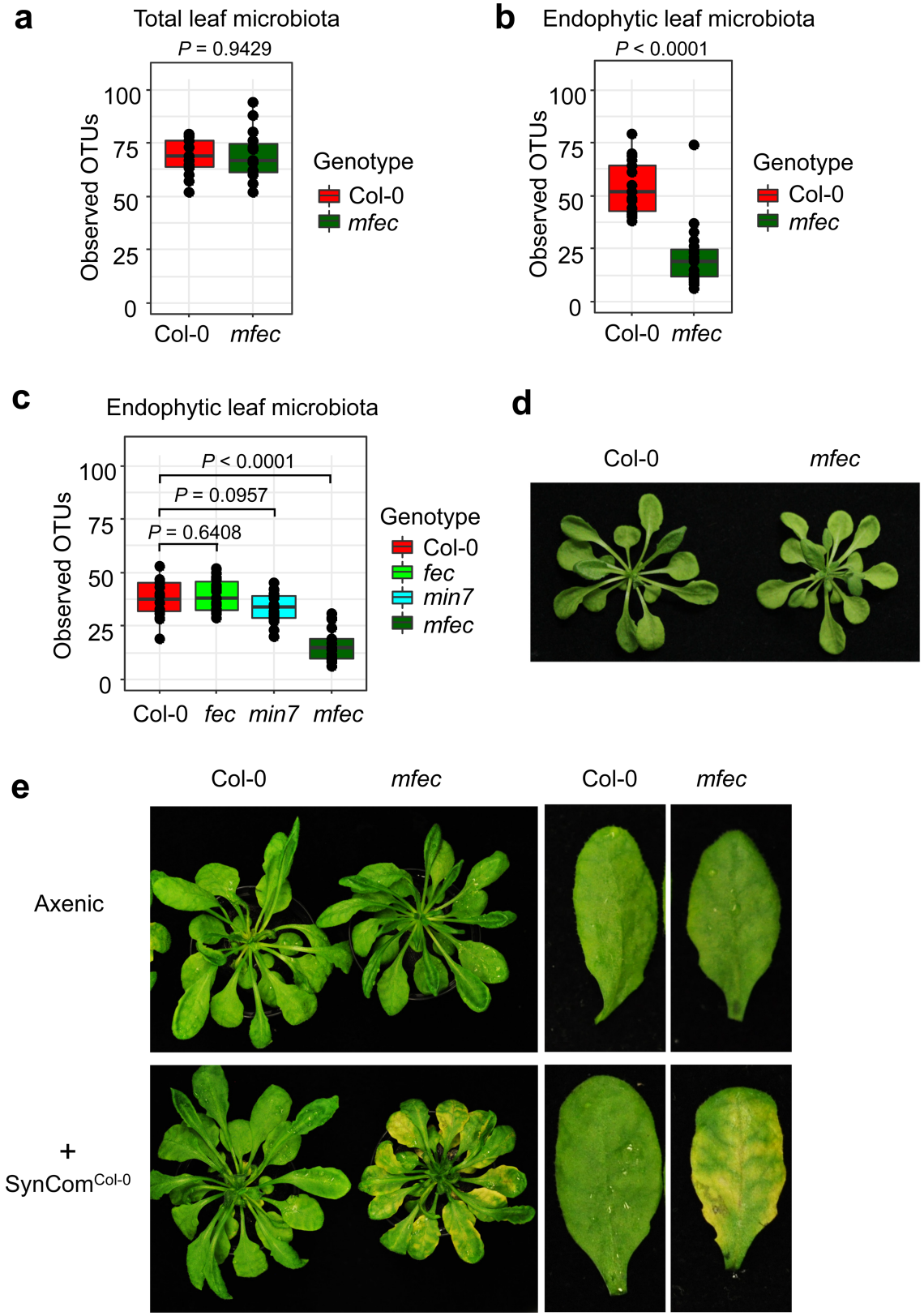
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2185-0.

Correspondence and requests for materials should be addressed to X.-F.X. or S.Y.H.

Peer review information *Nature* thanks Steven Lindow and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

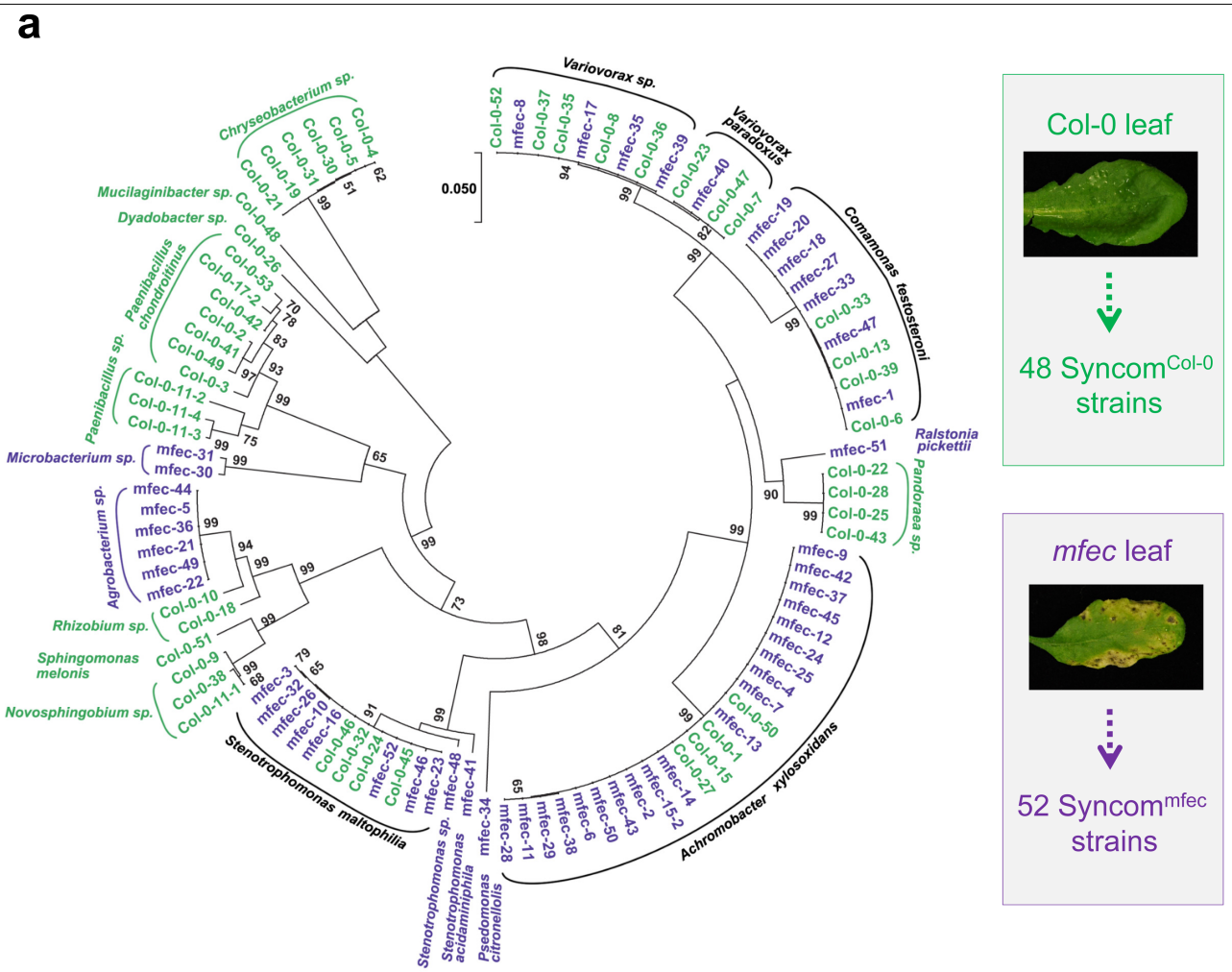


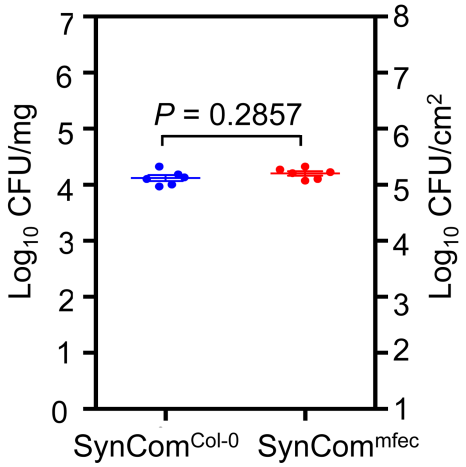
Extended Data Fig. 1 | Leaf and root appearance of soil-grown Col-0 and *mfec* plants. a, Leaf appearance of Col-0 and *mfec* plants grown in *Arabidopsis* mix soil (soil A) and Michigan State University (MSU) community agricultural soil (soil B; equal parts MSU community soil, medium vermiculate and perlite) or organic seed starter premium potting mix (Espoma) (soil C) for 6.5 weeks. Images were taken 5 days (soil A) or 11 days (soil B and soil C) after plants were shifted to high humidity (~95%). Representative leaf images are shown. **b**, Root appearance of Col-0 and *mfec* plants grown in *Arabidopsis* mix soil for five weeks and shifted to high humidity (~95%) for 5 days. Representative root images are shown. Experiments in **a**, **b**, were repeated three times with similar results.

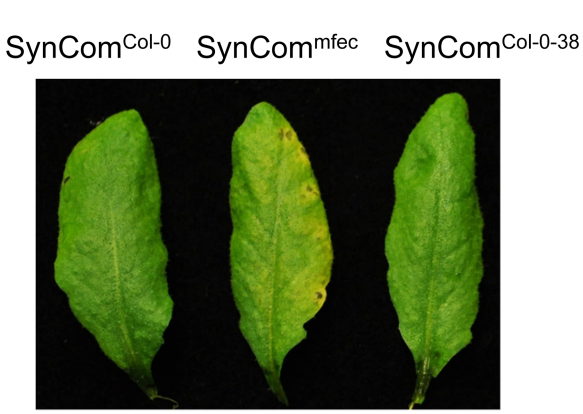


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Observed OTUs of total and endophytic leaf bacteria in different plant genotypes and requirement of microbiota for appearance of dysbiosis symptoms in *mfec* leaves. a, **b**, Observed OTUs of total (**a**) and endophytic leaf bacteria (**b**) in Col-0 and *mfec* plants, which were grown in *Arabidopsis* mix soil and shifted to high humidity for 5 days. **c**, Observed OTUs of endophytic leaf microbiota in Col-0, *fec*, *min7* and *mfec* plants supplemented with SynCom$^{Col\text{-}0}$. In box plots, the centre line represents the median, box edges show the 75th and 25th percentiles, and whiskers extend to 1.5× the interquartile range. Two-tailed Mann–Whitney *U*-test. $n = 15$ (Col-0) and $n = 15$ (*mfec*) biological replicates passing quality control for analysis of total leaf bacterial microbiota across 3 independent experiments; $n = 18$ (Col-0) and $n = 20$ (*mfec*) biological replicates passing quality control for analysis of leaf endophytic bacterial microbiota across 4 independent experiments. $n = 20$ (Col-0), $n = 19$ (*fec*), $n = 19$ (*min7*) and $n = 19$ (*mfec*) biological replicates passing quality control for analysis of leaf endophytic bacterial microbiota with SynCom$^{Col\text{-}0}$ across 4 independent experiments. **d**, Leaf appearance of Col-0 and *mfec* plants grown in sterile MS agar plates. Pictures were taken 5 days after shifting plates to high humidity (~95%). **e**, Leaf appearance of Col-0 and *mfec* plants grown in GnotoPots in the absence (axenic) or presence of SynCom$^{Col\text{-}0}$ for 6.5 weeks. Plants were then shifted to high humidity (~95%) for 10 days, before images were taken. Rosette leaf images are representative of at least four replicated experiments.

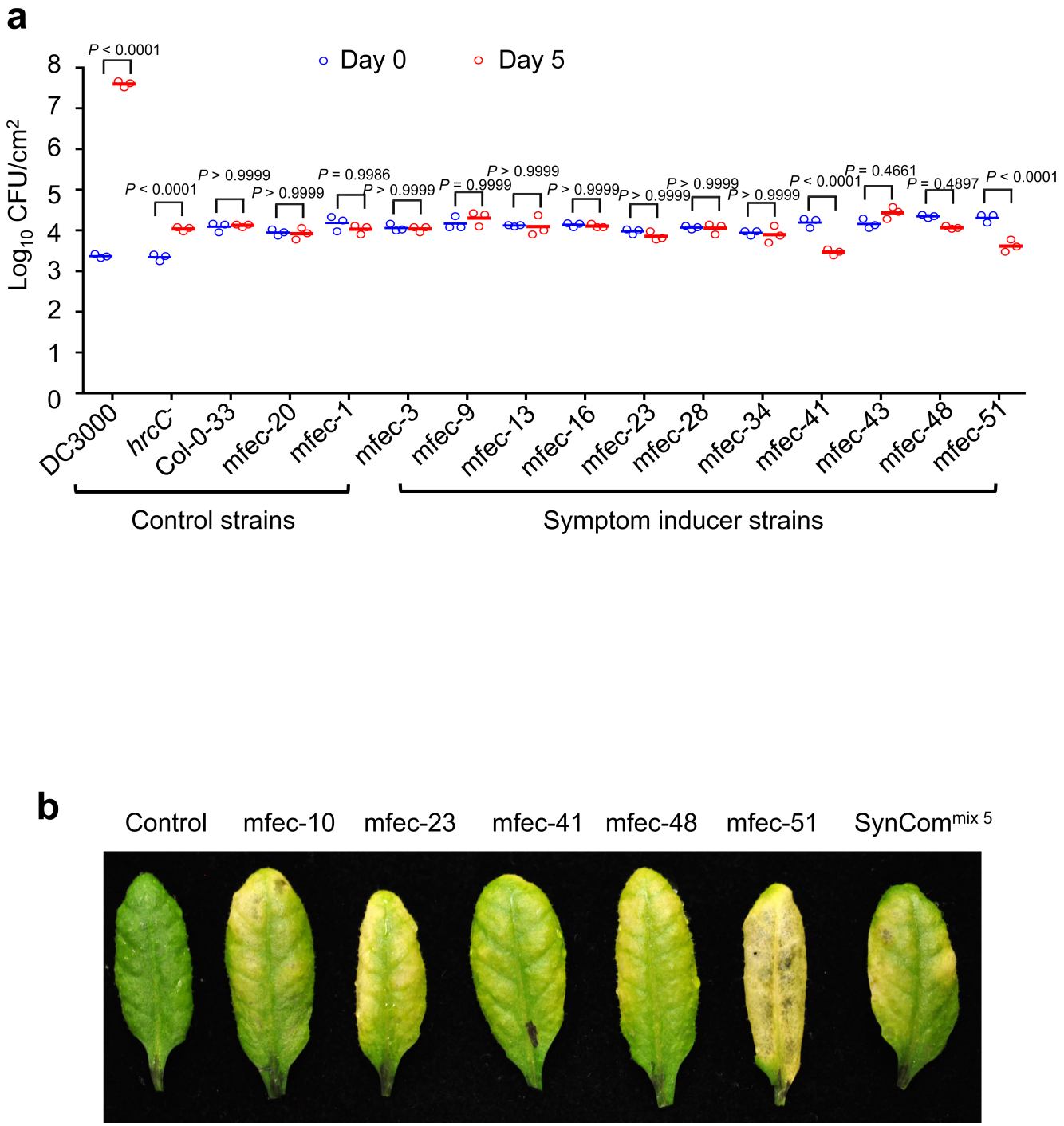




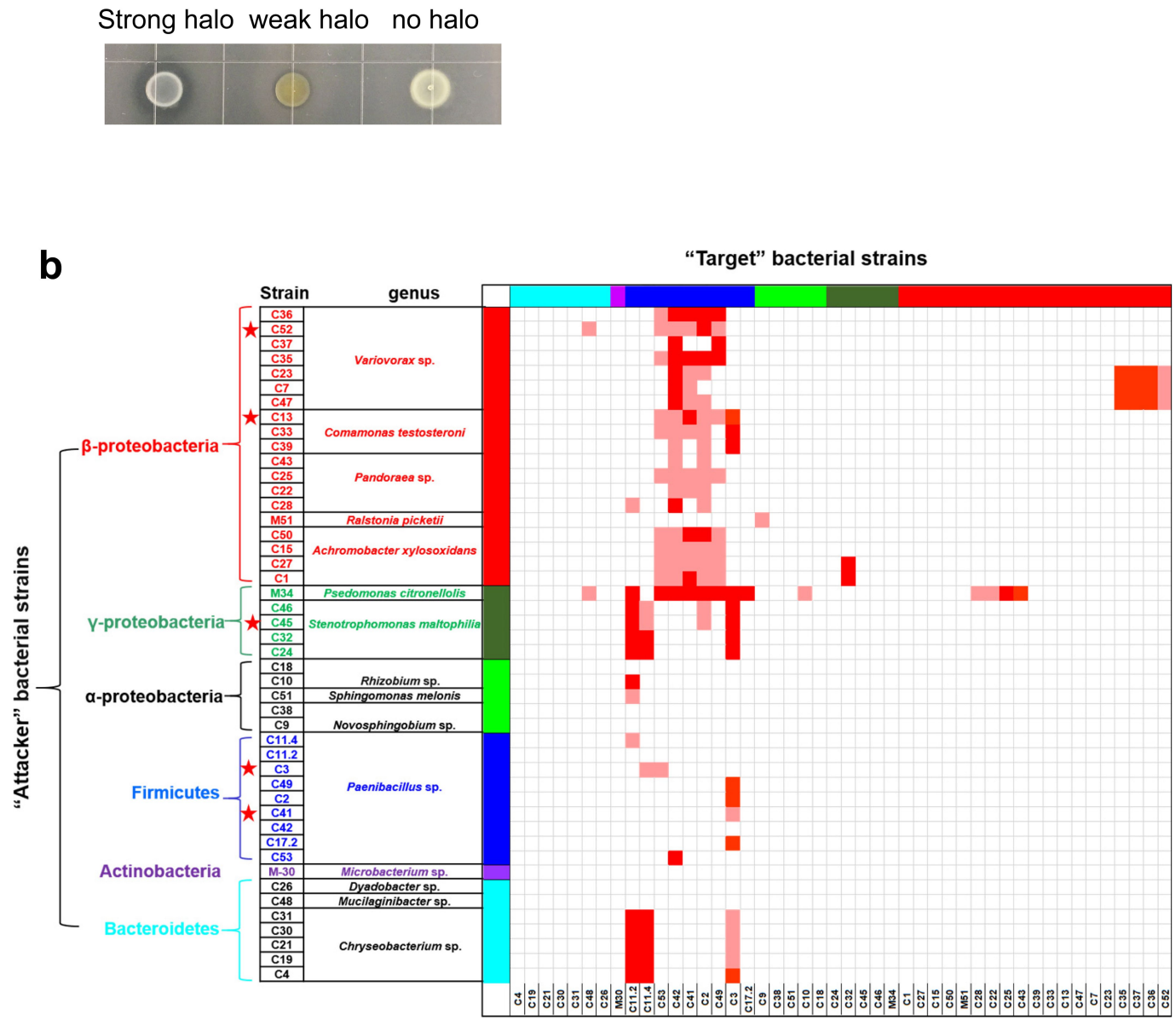


Extended Data Fig. 3 | See next page for caption.

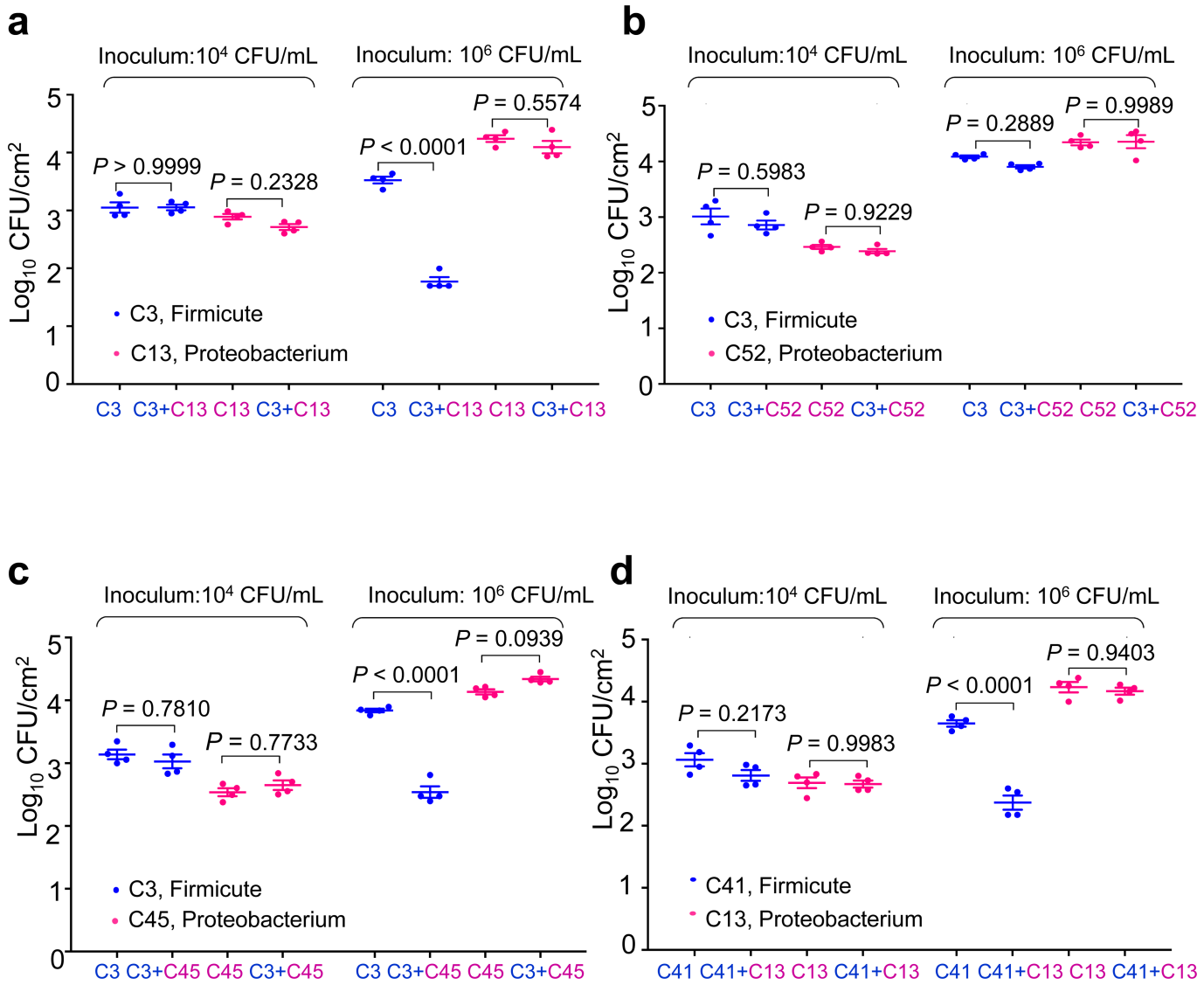
Extended Data Fig. 3 | A Maximum-likelihood phylogenetic tree for genome-sequenced bacterial isolates in SynCom$^{Col\text{-}0}$ and SynCom^{mfec}. a, Tree was constructed on the basis of the full-length 16S rRNA gene using MEGA7. A total of 100 bootstrap replicates were made, and bootstrap values are indicated at the branch points. Colours represent bacterial isolates from different plant genotypes: *mfec* mutant (purple); Col-0 (green). In total, 48 strains were derived from healthy Col-0 endophytic leaves and 52 strains were derived from *mfec* endophytic leaves displaying dysbiosis symptoms. **b**, Col-0 leaves were syringe-infiltrated with SynCom$^{Col\text{-}0}$ and SynCom^{mfec} at 1×10^7 CFU ml^{-1}; infiltrated plants were kept under ambient humidity for 1 h for water to evaporate. Bacterial populations were then determined after plant leaves returned to pre-infiltration appearance. Colony-forming units were normalized to tissue fresh weight (left) and leaf disk areas (right). Statistical significance was determined by two-tailed Mann–Whitney *U*-test. $n = 6$ biological replicates, data are mean ± s.e.m. Experiments were repeated three times with similar results. **c**, Col-0 plants were syringe-infiltrated with SynCom$^{Col\text{-}0}$, SynCom^{mfec} or SynCom$^{Col\text{-}0\text{-}38}$ (with 10 Firmicutes removed from SynCom$^{Col\text{-}0}$) at 1×10^7 CFU ml^{-1}. Inoculated plants were kept under high humidity (~ 95%), and leaf images were taken 7 days after infiltration. Experiments were repeated three times with similar results. Images are representative of leaves from four plants.



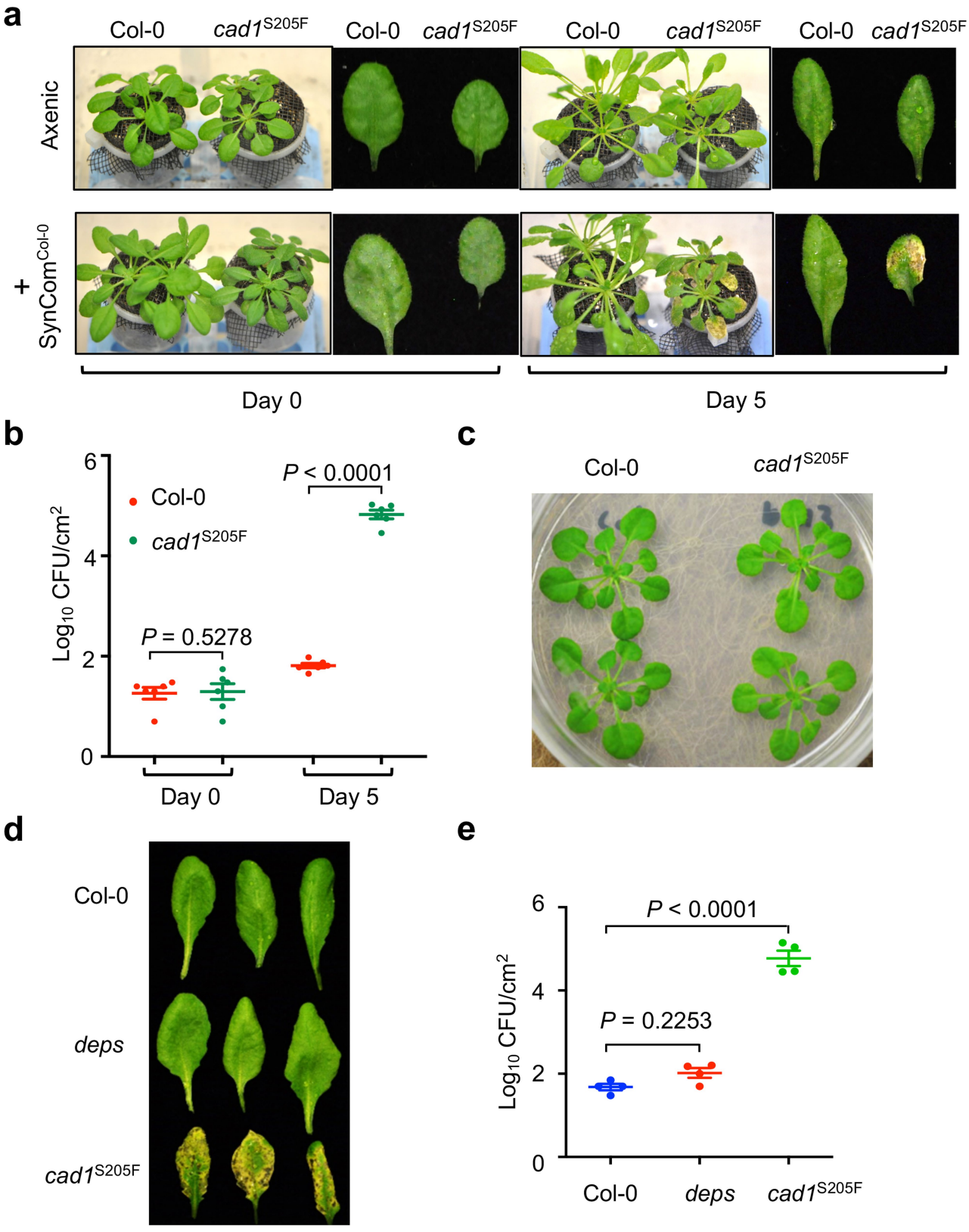
Extended Data Fig. 4 | Multiplication- and dysbiosis-symptom phenotypes of bacterial strains in Col-0 leaves. a, Population sizes (log_{10}CFU/cm^2 leaf area) of bacterial strains in Col-0 leaves on day 0 (1 h after leaf infiltration) and day 5 after leaf infiltration with each strain at 1×10^6 CFU ml^{-1}. The experiment was carried out at ~95% humidity. DC3000, *Pst* DC3000 (pathogenic on Col-0 plants); *hrcC*$^-$, a nonpathogenic mutant of DC3000 defective in type III secretion; Col-0-33, mfec-20 and mfec-1, control strains that do not induce dysbiosis symptoms (Supplementary Table 1); other mfec strains, induce dysbiosis symptoms (Supplementary Table 1). Statistical analysis was performed by two-way ANOVA with Tukey's test. $n = 3$ biological replicates, data are mean ± s.e.m. Experiments were repeated twice with similar results. **b**, Leaf dysbiosis symptoms 7 days after infiltration of leaves of 4.5-week-old Col-0 plants with indicated mfec strains or $SynCom^{mix\,5}$ at 1×10^7 CFU ml^{-1}. The experiment was carried out at ~95% humidity. $SynCom^{mix\,5}$ is a mix of mfec-10, mfec-23, mfec-41, mfec-48 or mfec-51 with equal OD_{600} values. Experiments were repeated three times with similar results.



Extended Data Fig. 5 | Binary inter-bacterial inhibition. a, Examples of inhibitory halos are labelled with strong, weak or no inhibition. **b**, Binary inhibition assays (2,116 combinations) on a R2A plate of 46 strains that represent all bacterial species identified in SynCom$^{Col\text{-}0}$ and SynCom^{mfec}. Target bacterial strains are presented along the horizontal axis, whereas attacker bacterial strains are listed vertically. A large or clear halo, indicative of strong binary inhibition, is represented by a red-filled cell; a small or less transparent halo, indicative of weaker binary inhibition, is represented by a pink-filled cell; the absence of halo is represented in white. Strains labelled with a star were used for the in planta binary inhibition assay in Extended Data Fig. 6. Experiments were repeated three times with similar results.

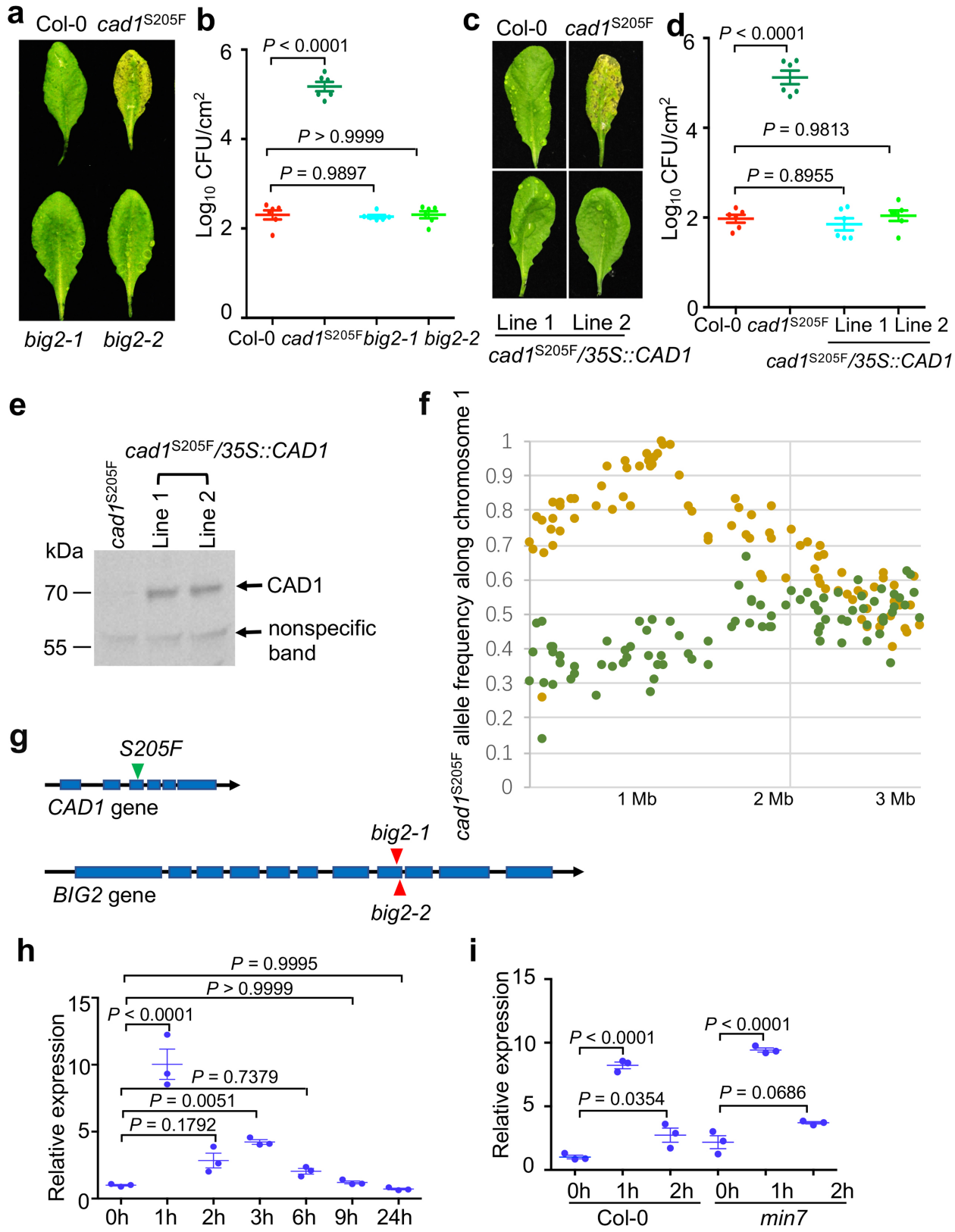


Extended Data Fig. 6 | In planta binary inhibition. a, In planta inhibition of Firmicutes by Proteobacteria strains that displayed a strong inhibitory effect in R2A agar plate assay. Leaves of Col-0 plants were syringe-infiltrated with *Paenibacillus chondroitinus* (C3; a Firmicutes) alone, *Comamonas testosteroni* (C13, a Proteobacterium) alone or C3 and C13 together at 1×10^4 CFU ml^{-1}, corresponding to approximately 1×10^2 CFU cm^{-2} leaf area; or 1×10^6 CFU ml^{-1}, corresponding to approximately 1×10^4 CFU cm^{-2} leaf area. After infiltration plants were maintained under high humidity (~95%) for 5 days before bacterial populations ($\log_{10}$CFU/cm^2 leaf area) were determined. **b**, Similar to **a**, but with a non-inhibitory binary interaction between strains C3 and *Variovorax* sp. C52 (a Proteobacterium). **c**, Leaves of Col-0 plants were syringe-infiltrated with *P. chondroitinus* (C3; a Firmicutes) alone, *Stenotrophomonas maltophilia* (C45, a Proteobacterium) alone or C3 and C45 together at 1×10^4 CFU ml^{-1} or 1×10^6 CFU ml^{-1}. **d**, Leaves of Col-0 plants were syringe-infiltrated with *P. chondroitinus* (C41; a Firmicutes) alone, *C. testosteroni* (C13, a Proteobacterium) alone or C41 and C13 together at 1×10^4 CFU ml^{-1} or 1×10^6 CFU ml^{-1}. After infiltration, plants were maintained under high humidity (~95%) for 5 days before bacterial populations were determined. One-way ANOVA with Tukey's test. $n = 4$ biological replicates, data are mean ± s.e.m. Experiments were repeated three times with similar results.



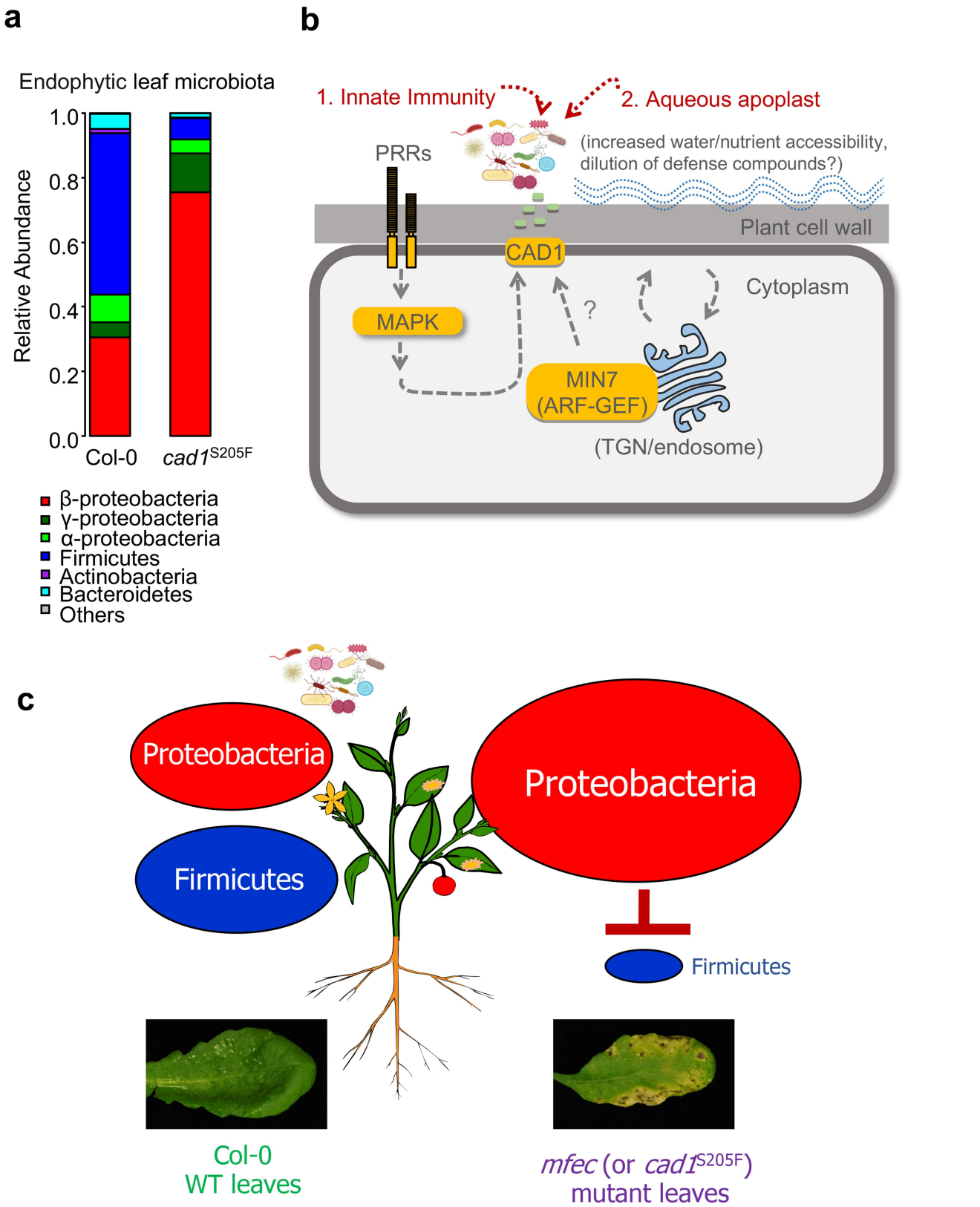
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Appearance and bacterial populations in Col-0 and *cad1*^{S205F} plants before and after shift to 95% humidity. a, Leaf appearance of 5-week-old Col-0 and *cad1*^{S205F} plants grown in the absence (axenic) or presence of SynCom$^{Col\text{-}0}$ in the FlowPot gnotobiotic system (see Methods). Images were taken before (day 0) and 5 days after plants were shifted to high humidity (~95%). **b**, Levels of endophytic bacterial community (log_{10}CFU/cm^2 leaf area) in the presence of SynCom$^{Col\text{-}0}$ in the FlowPot gnotobiotic system. One-way ANOVA with Tukey's test. Data are mean ± s.e.m., $n = 6$ biological replicates. Experiments were repeated three times with similar results. **c**, Leaf appearance of Col-0 and *cad1*^{S205F} plants grown in enclosed sterile LS agar plates for 4 weeks. **d**, Leaf appearance of 5-week-old Col-0, *deps* and *cad1*^{S205F} plants grown in *Arabidopsis* mix 5 days after plants were shifted to ~95% relative humidity. **e**, Levels of endophytic leaf microbiota (log_{10}CFU/cm^2 leaf area) in 5-week-old Col-0, *deps* and *cad1*^{S205F} plants 5 days after plants were exposed to high humidity (~95%). One-way ANOVA with Tukey's test. Data are mean ± s.e.m., $n = 4$ biological replicates. Experiments were repeated three times with similar results.

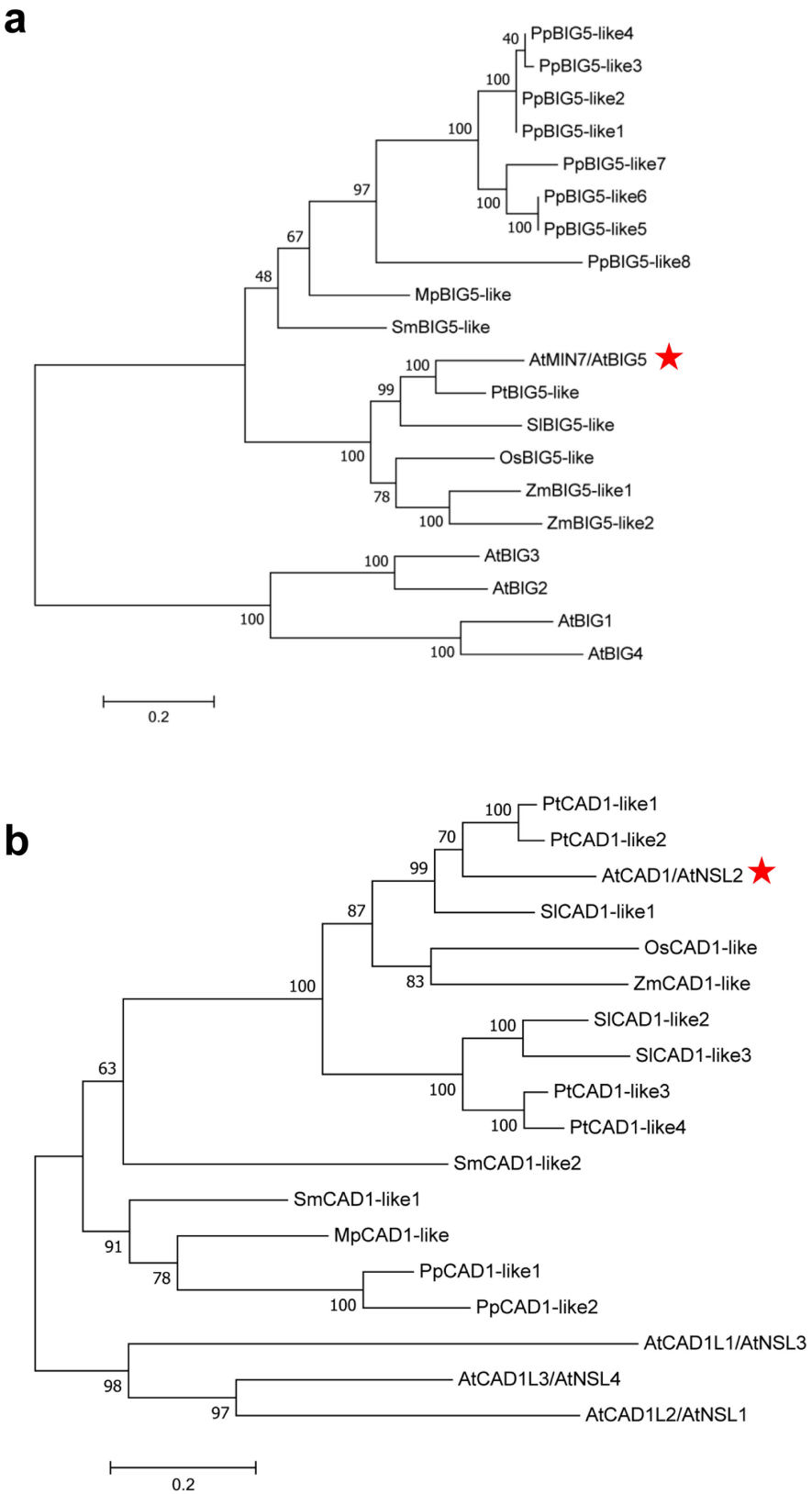


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Identification of a *cad1* mutation responsible for dysbiosis in the *cad1* mutant. a, Leaf appearance of 4.5-week-old Col-0, *cad1*^{S205F} and *big2* plants grown in redi-earth potting soil. Images were taken at day 5 after plants were shifted to 95% humidity. **b**, Bacterial populations (log_{10}CFU/cm^2 leaf area) of the endophytic bacterial community. One-way ANOVA with Tukey's test. Data are mean ± s.e.m., $n = 6$ biological replicates. Experiments were repeated three times with similar results. Two independent T-DNA insertion lines of *BIG2* were analysed with similar results (*big2-1*, SALK_033446 and *big2-2*, SALK_016558). **c**–**e**, Appearance (**c**) and endophytic bacterial populations (**d**; log_{10}CFU/cm^2 leaf area) in Col-0, *cad1*^{S205F} and *cad1*^{S205F}*/35S::CAD1* plants at day 5 after plants were shifted to high humidity. Plants were grown in redi-earth potting soil for 4.5 weeks before they were shifted to high humidity. One-way ANOVA with Tukey's test. Data are mean ± s.e.m., $n = 6$ biological replicates. Experiments were repeated three times with similar results. **e**, Two independent different complementation lines (*cad1*^{S205F}*/35S::CAD1* line 1 and *cad1*^{S205F}*/35S::CAD1* line 2) were analysed with similar results and protein levels were confirmed by western blot with the CAD1 antibody. Uncropped gel image is shown in Supplementary Fig. 2. **f**, *cad1*^{S205F} genomic mapping. Green and brown dots indicate wild type-like and *cad1*^{S205F}-like allele frequencies, respectively (detailed information in Supplementary Table 6). **g**, Schematic of mutations in *big2* and *cad1*^{S205F} mutants. **h**, **i**, Quantitative PCR analyses of *CAD1* transcript in Col-0 (**h**) and *min7* (**i**) plants grown in *Arabidopsis* mix soil. Five-week-old Col-0 and *min7* leaves were infiltrated with 1 µM flg22 and collected at the indicated time points. Transcript levels were normalized to that of the *PP2AA3* gene. One-way ANOVA with Tukey's test. Data are mean ± s.e.m., $n = 3$ biological replicates. Experiments were repeated three times with similar results.



Extended Data Fig. 9 | A model for plant control of endophytic phyllosphere microbiota. a, The 16S rRNA gene-sequence profiles of endophytic leaf bacteria in Col-0 and *cad1*^{S205F} plants supplemented with SynCom^{Col-0}. Data presentation and statistical analysis as in Fig. 1d. $n = 20$ (Col-0) and $n = 20$ (*cad1*^{S205F}) biological replicates. **b**, A simplified diagram depicting pattern-triggered immune signalling, MIN7 and CAD1 as three components of a putative genetic framework for controlling endophytic bacterial microbiota, which live outside a plant cell. MIN7 has previously been shown to be involved in regulating callose deposition[51,52] and aqueous microenvironment in the leaf apoplast (that is, extracellular space)[1]. **c**, Large shifts in the level and composition of endophytic leaf microbiota in wild-type Col-0 versus *mfec* (or *cad1*^{S205F}) leaves in part via competition between Proteobacteria and Firmicutes. Some components in **b** and **c** were drawn using tools in biorender.com.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Phylogenetic trees of protein sequences of MIN7 and CAD1 homologues from different plant species. a, **b**, Protein sequences of *A. thaliana* AtMIN7 (also known as AtBIG5) (AT3G43300.1) (**a**) and AtCAD1 (also known as AtNSL2) (AT1G29690.1) (**b**) were used for comparisons by Blast search against the proteome of *Arabidopsis* and seven other plant species (https://phytozome.jgi.doe.gov/). Homologues with *E* values lower than E^{100} were selected to generate phylogenetic trees across taxa, and only homologues specific to the AtMIN7 or AtCAD1 clade were presented with selected proteins from *Arabidopsis* as outgroups. Bootstrap values were obtained from 1,000 replicates using the maximum-likelihood algorithm using MEGA7. The scale bar represents 0.2 substitutions per amino acid site. The genes are listed in Supplementary Table 7. AtMIN7 and AtCAD1 are highlighted with red stars. Abbreviations: BIG, BREFELDIN A-INHIBITED GUANINE NUCLEOTIDE-EXCHANGE PROTEIN; NSL, NECROTIC SPOTTED LESIONS; At, *A. thaliana*; Mp, *Marchantia polymorpha*; Os, *Oryza sativa*; Pp, *Physcomitrella patens*; Pt, *Populus trichocarpa*; Sm, *Selaginella meollendorffii*; Sl, *Solanum lycopersicum*; Zm, *Zostera marina*.

nature research

Corresponding author(s): Sheng Yang He and Xiu-Fang Xin

Last updated by author(s): Feb 10, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Not applicable.

Data analysis	Description of published software used for data analysis is provided and cited in the Methods section (pages 9-14) and figure legends. These include: Canu version 1.7: https://canu.readthedocs.io/en/latest/ Sickle version 1.33: https://github.com/najoshi/sickle Pilon version 1.22: https://github.com/broadinstitute/pilon Gtdbtk version 0.1.3: https://github.com/Ecogenomics/GTDBTk Mothur version v.1.34.2: https://www.mothur.org/ MEGA7: https://www.megasoftware.net/ Easy Leaf Area: https://github.com/heaslon/Easy-Leaf-Area QIIME 2 Core 2018.11 distribution: https://qiime2.org/ Cutadapt: https://github.com/qiime2/q2-cutadapt DADA2: http://benjjneb.github.io/dada2/ edgeR package: https://bioconductor.org/packages/release/bioc/html/edgeR.html Trimmomatic (v0.33): http://www.usadellab.org/cms/?page=trimmomatic bowtie2 (v2.3.1): http://bowtie-bio.sourceforge.net/bowtie2/index.shtml SAMtools (v1.5): http://www.htslib.org/ picardTools (v1.89): https://broadinstitute.github.io/picard/ bcftools (v1.2): http://www.htslib.org/ SHOREmap v3.0: http://bioinfo.mpipz.mpg.de/shoremap/ Scripts Scripts used in the microbiota analyses are available at https://github.com/godlovexiaolin/A-genetic-network-for-host-control-of-phyllosphere-microbiota-for-plant-health.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw 16S rRNA gene sequences from this project are available in the SRA (Sequence Read Archive) database under the BioProject PRJNA554246, accession no. SAMN12259846- SAMN12260169. Bacterial genome data are available in the SRA database under the BioProject PRJNA555902.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size and the results of statistical analyses are described in the relevant figure legends. Sample size was determined based on experimental trials and with consideration of previous publications on similar experiments to allow for confident statistical analyses. No statistical methods were used to predetermine sample sizes.
Data exclusions	No data that pass quality control were excluded from analysis.
Replication	The number of replication for each experiment (at least two repeats, but mostly three times) is described in the relevant figure legends. Results were reproducible in all repeats with the same trend.
Randomization	Plants of different genotypes (Col-0, mfec, and cad1) were grown side by side to minimize unexpected environmental variations during growth and experimentation. Leaf samples of similar ages were collected and assessed randomly for each genotype.
Blinding	Researchers were not blinded to allocation during experiments and outcome assessment. This is in part because different plant genotypes under study (Col-0, mfec, and cad1) exhibit very distinct phenotypes visually; blinding was not possible. Routine practices included more than one author observing/assessing phenotypes, whenever possible.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: A Guinea Pig antibody to the Arabidopsis CAD1 protein; custom-made at Cocalico Biologicals, Inc (animal number: Guinea pig MSU-GP-19 and 20). Dilution at 1:5,000.

Validation: CAD1 antibody was validated by immunoblot of the CAD1 protein present in wild-type and absence in the cad1 mutant Arabidopsis.

Action of a minimal contractile bactericidal nanomachine

https://doi.org/10.1038/s41586-020-2186-z

Received: 26 April 2018

Accepted: 6 February 2020

Published online: 15 April 2020

Check for updates

Peng Ge[1,2,8], Dean Scholl[3,8], Nikolai S. Prokhorov[4], Jaycob Avaylon[2,5], Mikhail M. Shneider[6], Christopher Browning[7], Sergey A. Buth[4], Michel Plattner[4], Urmi Chakraborty[3], Ke Ding[1,2], Petr G. Leiman[4], Jeff F. Miller[1,2]✉ & Z. Hong Zhou[1,2]✉

R-type bacteriocins are minimal contractile nanomachines that hold promise as precision antibiotics[1–4]. Each bactericidal complex uses a collar to bridge a hollow tube with a contractile sheath loaded in a metastable state by a baseplate scaffold[1,2]. Fine-tuning of such nucleic acid-free protein machines for precision medicine calls for an atomic description of the entire complex and contraction mechanism, which is not available from baseplate structures of the (DNA-containing) T4 bacteriophage[5]. Here we report the atomic model of the complete R2 pyocin in its pre-contraction and post-contraction states, each containing 384 subunits of 11 unique atomic models of 10 gene products. Comparison of these structures suggests the following sequence of events during pyocin contraction: tail fibres trigger lateral dissociation of baseplate triplexes; the dissociation then initiates a cascade of events leading to sheath contraction; and this contraction converts chemical energy into mechanical force to drive the iron-tipped tube across the bacterial cell surface, killing the bacterium.

Contractile nanotube-based machines are widespread in the bacterial domain, functioning to penetrate cell membranes to deliver payloads of proteins or DNA, or to create channels through which ions translocate[6]. Contractile type VI secretion systems (T6SSs) inject proteins into eukaryotic or bacterial cells to facilitate pathogenesis or to kill competing organisms, respectively[7–13]. Phage tail-like bacteriocins, exemplified by the R-type pyocins produced by *Pseudomonas aeruginosa*, employ the same contractility to kill competing bacteria by dissipating their membrane potential[1,2,14]. Myovirus bacteriophages, such as P2 and T4, use a similar contractile mechanism to translocate DNA into bacterial cells[5,9,15–22]. These contractile nanomachines employ a spring-loaded sheath–tube assembly to penetrate target cell surfaces[11]. This process is accompanied by massive structural transformations involving contraction of the sheath triggered by the baseplate, which has been visualized by both single-particle cryo-electron microscopy (cryo-EM)[1,7,10,23,24] and cellular cryo-electron tomography[25]. Although energy storage is mostly similar in phage, pyocin and the T6SS, the triggering mechanism may be different, owing to the presence of additional cell membranes in the case of the T6SS. By contrast, some other delivery systems, such as the T3SS and T4SS[26,27], carry no stored energy for penetration or baseplate-like structure for triggering.

An engineered T4 tube–baseplate complex reported previously[5] shows the static structures of baseplate proteins but lacks the sheath to inform how it receives the contraction signal. More recently, cryo-EM structures of insecticidal contractile toxin-delivery systems from *Photorhabdus asymbiotica* (*Photorhabdus* virulence cassette[23]) and *Serratia entomophila* (antifeeding prophage[24]) showcase the widespread existence of such phage tail-like contractile systems in nature. The relative simplicity and ease of engineering[3,4] of R-type pyocins make them ideal model systems for studying contractile structures. Our previous helical reconstructions of the pyocin R2 sheath and tube in pre-contraction and post-contraction states[1] revealed how energy for contraction is stored and released by shape and charge complementarity. However, the lack of atomic detail on the baseplate and collar precluded understanding of the molecular trigger that initiates sheath contraction, or how the resulting structure is stabilized to facilitate killing.

Here we report the atomic models of the pyocin collar and baseplate derived from single-particle cryo-EM and X-ray crystallography in their pre-contraction and post-contraction states. By comparing these structures, we derived and tested a model for the action of a bactericidal nanomachine that couples specific recognition of target cells with deployment of a generic mechanism of killing, providing insights that are crucial for exploiting these structures as precision antimicrobials[2,28,29].

Overall structures of pyocin R2

We imaged pyocin R2 by cryo-EM (Fig. 1a, b). Under the conditions used for sample isolation, both pre-contracted and post-contracted particles are present in the same preparation[1]. Each pyocin R2 complex consists of three structurally and functionally distinct components: the collar, the trunk and the baseplate with tail fibres. In the post-contraction state, the sheath layer of the trunk contracts by 70%, leaving the central tube exposed and readily visible in the cryo-EM images (Fig. 1a, b). We determined the cryo-EM structures of the pre-contracted pyocin R2 collar, trunk and baseplate regions

[1]Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles (UCLA), Los Angeles, CA, USA. [2]The California NanoSystems Institute (CNSI), University of California, Los Angeles (UCLA), Los Angeles, CA, USA. [3]Pylum Biosciences, South San Francisco, CA, USA. [4]University of Texas Medical Branch, Department of Biochemistry and Molecular Biology, Sealy Center for Structural Biology and Molecular Biophysics, Galveston, TX, USA. [5]Department of Chemistry and Biochemistry, University of California, Los Angeles (UCLA), Los Angeles, CA, USA. [6]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Laboratory of Molecular Bioengineering, Moscow, Russia. [7]Vertex Pharmaceuticals (Europe) Ltd, Abingdon, UK. [8]These authors contributed equally: Peng Ge, Dean Scholl. ✉e-mail: jfmiller@UCLA.edu; Hong.Zhou@UCLA.edu

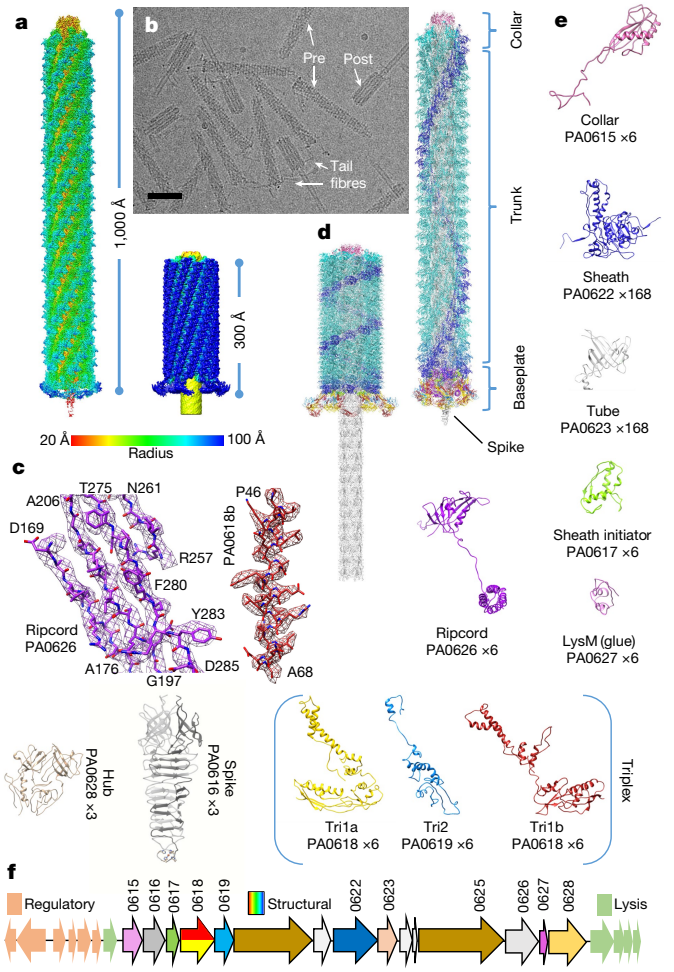


Fig. 1 | Cryo-EM and overall structure of pyocin in pre-contracted and post-contracted states. a, Shaded surface representation of the cryo-EM reconstructions, coloured according to cylindrical radii as shown in the colour bar. **b**, A representative cryo-EM micrograph. Scale bar, 300 Å. **c**, Regions of the cryo-EM density map (mesh) superimposed with atomic models (sticks) demonstrating the agreement between the observed and modelled amino acid side chains. **d**, Atomic models for pyocin in both the pre-contracted (right) and the post-contracted (left) states. **e**, **f**, Ribbon diagrams of individual subunits of pyocin in the pre-contracted state (**e**) shown along with their corresponding gene loci (**f**). See the 3D rendition in Supplementary Videos 1–3.

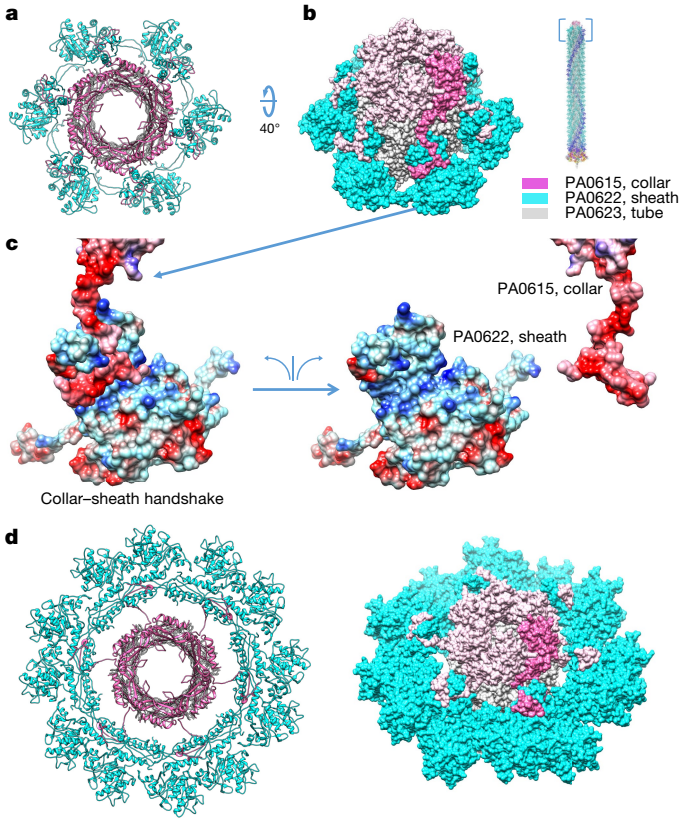


Fig. 2 | Architecture of the collar. a, Top view ribbon diagram of the collar (pink), the outer sheath (cyan) and the inner tube (grey). **b**, Space filling model of the collar–sheath–tube region. **c**, Electrostatic surface model of the collar–sheath handshake with views of the interacting charged surfaces (positive (blue), negative (red) and neutral (same as **b**)). **d**, Ribbon (left) and space filling (right) diagrams of the post-contraction collar–sheath–tube region similar to **a** and **b**, respectively.

separately and made a montage model by computationally stitching the three parts together (Fig. 1a, b and Supplementary Video 1). Substructures were determined at resolutions of 3.8 Å for the collar, 2.9 Å for the trunk and 3.5 Å for the baseplate (Extended Data Figs. 1, 2 and Extended Data Table 1). Although the average resolution of the baseplate reaches 3.5 Å (Fig. 1c and Supplementary Video 2), the resolution of the peripheral regions is lower, possibly owing to blurring from Brownian motion of the connected tail fibres (these regions are modelled according to both their cryo-EM density and our crystal structure of an engineered protein that represents them (Protein Data Bank (PDB) ID: 5CES) (Extended Data Table 2)). Using the same strategy, we determined the structure of the post-contracted pyocin at intermediate resolutions (Fig. 1a, right, and Supplementary Video 3). By matching amino acid side chains that are visible in our cryo-EM structures, we assigned a single gene product, PA0615, to the collar; two gene products, PA0622 and PA0623, to the trunk (PA0622 sheath and PA0623 tube); and seven gene products, PA0616–PA0619 and PA0626–PA0628, to the various components of the baseplate. We then built an atomic model of the complete pyocin, excluding the tail fibre and tape measure proteins (PA0620 and PA0625, respectively) (Fig. 1d–f, Extended Data Table 3 and Supplementary Video 4).

Collar tethers tube to the contracted sheath

The collar is a hexamer formed by the gene product PA0615 (Fig. 2a, b). Each collar monomer has a simple structure with two domains, one globular domain and one β-hairpin domain, joined by an extended loop (Figs. 1e and 2b). The structure of the globular domain is similar to that of the inner tube protein. The collar extends the tube and tethers it to the sheath, thus preventing the tube from dissociating from the sheath after contraction as demonstrated in the *Photorhabdus* virulence cassette[23]. By capping the top of the tube and augmenting the handshake β-sheet of the sheath subunit below with a β-hairpin (similar to the sheath–sheath handshake[1,7,10]), it provides mechanical stability to the junction after the downward pull of the sheath against the tube (Fig. 2a, b).

Specifically, the handshake domain of each sheath subunit is augmented by two β-strands that are donated by subunits from a disc that is closer to the collar. Thus, each subunit of the last disc has a β-sheet with two vacant β-strands. The β-hairpin domain of the collar protein completes this β-sheet. In addition, the hydrogen bond interactions in this augmentation are reinforced by charge–charge interactions: the loop and the β-hairpin domain of the collar are mainly negative, binding to a groove in the sheath that is mainly positive (Fig. 2c), unlike the very

hydrophobic nature of the sheath–sheath interaction. In this configuration, the collar hexamer joins the tube and the sheath at the very top.

We also determined the structure of the post-contraction collar region at an average resolution of 3.5 Å (Extended Data Fig. 1). The most dramatic structural change is that the diameter of the outer sheath increases from 180 Å to 240 Å, which results in the dislocation of the sheath from the tube (Fig. 2d). This dislocation appears to reduce the structural rigidity of the collar protein linker and to increase the local mobility of the tube, as the averaged resolution of the tube portion in the reconstruction is only 7 Å and the linker has become invisible. Nonetheless, the rest of the structure, including the sheath protein and the hairpin domain of the collar, is of sufficient resolution for building atomic models. The hairpin domain of the collar undergoes a slight conformational change during contraction, but the globular domain remains structurally unchanged.

The trunk and the sheath initiator

Since we determined the structures of the bottommost part of the trunk and the baseplate without imposing helical symmetry, we can now resolve the interface regions of both components. Six subunits of the PA0623 tube protein form a hexameric ring with a 24-stranded β-barrel, 28 of which stack into the central tube of each pyocin. Handshake interactions with long N-terminal and C-terminal extension arms interweave PA0622 sheath protein subunits, maintaining the connectivity of the sheath during contraction (Extended Data Fig. 3). Thus, our new non-helical structure confirms our previous observations based on helical reconstructions of the trunk portion of pyocin from images recorded on film[1].

Sheath discs that approach the baseplate break from the shared helical symmetry of the trunk with a slightly but gradually increasing rotation per disc, reaching an extra 4.4° in the bottommost disc (Extended Data Fig. 4). The handshake β-sheet is the same across the entire sheath, except for the bottommost disc where the sheath subunit arms are incorporated into the PA0617 sheath initiator protein (Extended Data Fig. 3). The arm of the sheath protein augments a β-sheet of the sheath initiator protein in the same manner as in the rest of the sheath. Thus, the β-sheet augmentation mechanism allows the sheath to be linked via multiple polypeptide chains to both the collar and the baseplate (see the three different types of handshaking augmentation in Extended Data Fig. 3).

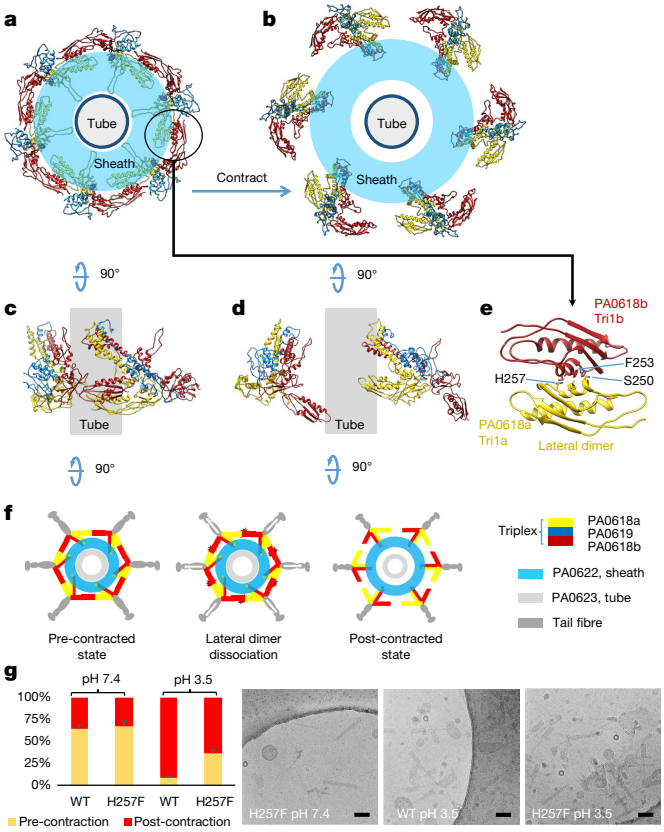


Fig. 3 | Triplex expansion and lateral dimer dissociation. a–**e**, Ribbon diagrams of triplexes forming an iris ring in the pre-contracted state (**a**), an expanded iris in the post-contracted state (**b**), side view of two adjacent triplexes in the pre-contracted state (**c**) and in the post-contracted state (**d**), and a lateral dimer of PA0618 (**e**). **f**, Schematic of the iris ring expansion as a result of the tail fibre actuation. **g**, The PA0618 H257F mutant. Left, percentages of pre-contraction pyocins in the purified WT and the H257F mutant under cryo-EM at neutral and acidic pH (counts: pre-contraction/all: pH 7.4 WT: 185/289, H257F: 118/175; pH 3.5 WT: 46/530, H257F: 64/178). Error bars represent standard deviations. Right, the representative cryo-EM image for each relevant condition. Scale bars, 300 Å.

Baseplate: triggering triplex and ripcord

Our reconstruction shows that the pyocin baseplate is composed of eight different protein subunits: ripcord (PA0626), triplex (two copies of PA0618 (Tri1a and Tri1b) and one copy of PA0619 (Tri2)), sheath initiator (PA0617), glue (PA0627), hub (PA0628) and spike (PA016) (Fig. 1e and Extended Data Fig. 5). PA0626 forms the centrepiece of the baseplate (Extended Data Fig. 5a): the central spike complex is 'below', the tube–sheath and the trunk that they form is 'above', and the rest of the baseplate (triplex, sheath initiator and glue) surrounds it. The central spike complex is composed of the trimeric or pseudo-hexameric PA0628 hub (Extended Data Fig. 5b) and the trimeric PA0616 central spike inserted into it (Extended Data Fig. 5c). Our crystal structure of the PA0616 spike, a homologue of the P2 phage gpV[30], shows that it carries a ferric ion at the very tip. The ferric ion is coordinated by a triplet of H×H double histidine motifs, a structure that stabilizes the tip for membrane penetration (Extended Data Fig. 5c).

Each PA0626 monomer has a C-terminal β-sheet domain (amino acids 136–286) and an N-terminal helix-rich domain (amino acids 1–112) joined by an extended linker (amino acids 113–135), which is an unusual domain organization reminiscent of a ripcord (Extended Data Fig. 5a). We reason that this setup may comprise a form of an activation energy barrier for triggering and may also be crucial for assembly: the C-terminal domains of six PA0626 subunits form a hexameric disc, which resembles that of the tube proteins and extends the tube (Extended Data Fig. 5a). This disc is integral to the tube–PA0626–hub–spike assembly that penetrates the target as it connects the tube to the hub. The N-terminal domains of PA0626, a four-helical bundle structure, are localized at a higher cylindrical radius inside the baseplate (Extended Data Fig. 5d). Indeed, when we introduce a tobacco etch virus (TEV) protease cleavage site in the linker of the ripcord, the mutant pyocin (626TEV) shows wild-type (WT)-like assembly without TEV protease co-expressed and no assembly with TEV protease co-expressed (Extended Data Fig. 6). We also found that this mutant (626TEV) and another deletion mutant (626ΔWL) have lower activation energy for triggering compared to the WT (Extended Data Fig. 6).

Two copies of PA0618 and one copy of PA0619 proteins form a $(PA0618)_2$–PA0619 heterotrimeric triplex by attaching two copies of PA0618 (Tri1a/Tri1b) to either side of PA0619 (Tri2), resulting in two distinct conformations of PA0618 (Extended Data Fig. 7), which accept the PA0626 N-terminal four-helical bundle domain (Extended Data Fig. 8). At its top, this triplex also binds to the sheath initiator PA0617 protein. A small protein with a LysM fold—PA0627—binds to a side of the triplex and glues PA0617, PA0618b and PA0619 together (Extended Data Fig. 5e). The $(PA0618)_2$–PA0619 triplexes are joined into an iris-like

structure by lateral dimers of the C-terminal domains of PA0618 (Fig. 3a and Extended Data Fig. 7e, f). PA0619 attaches to the tail fibre, allowing it to receive a triggering signal from it (Extended Data Fig. 5f).

By comparing the pre-contraction and post-contraction baseplate models, we found that the iris ring[5], joined by lateral dimerization of PA0618, breaks apart after contraction of the pyocin (Figs. 1d and 3). In this way, the baseplate in the post-contraction state splits into a hexagram shape. The entire complex of the sheath initiator, glue and triplex travels in a rigid-body movement as each sheath subunit does, widening the baseplate to 320 Å (Fig. 3b). The tail fibre still binds to the triplex upon contraction, as its densities are still visibly connected to the triplex, although at low resolution.

Compared to the bacteriophage T4 baseplate[5], pyocin baseplate proteins are minimalistic and bear interesting differences. PA0618 and PA0619 both lack sizeable insertion domains that are required for building the considerably larger T4 baseplate. Instead, these insertions are replaced by loops (Extended Data Fig. 9), although their function as struts that connect the hub and the baseplate is preserved. PA0617 and PA0627 are also minimalistic compared to T4 gp25 and gp53, respectively, simply retaining the core motifs[5] (Extended Data Fig. 9). Furthermore, the pyocin sheath protein lacks domains 3 and 4 of the T4 sheath protein, which add to the energy release of contraction and interact with long tail fibres, respectively[16,17].

Bactericidal action and application

On the basis of our atomic models of R2 pyocin in pre-contracted and post-contracted states, we constructed a minimal contractile machine that includes just 12 stacks of the sheath and rendered a morph movie between the two states to illustrate a possible pathway of action of such a contractile nanomachine (Supplementary Video 5). Notably, because the sheath changes helicity during contraction, the tube undergoes a rotational movement during its power stroke, which may facilitate the spike's penetration of the target cell surface.

The biological function of R-type pyocins is to kill competing bacteria and they do so with an extraordinary efficiency that approaches single-shot killing[4,31]. To achieve this, it is necessary to actuate only in the right place and time. Upon recognition by tail fibre receptor-binding proteins of specific ligands on a target cell, the increased free energy would cause displacement of interacting surfaces within lateral dimers of PA0618, and we suggest that shearing forces that are transduced through tail fibres trigger the baseplate to initiate contraction (Fig. 4). Shearing force encountered by particles during purification may bear similar characteristics. This force drives the baseplate to transit to a larger diameter, breaking the lateral dimers of PA0618 Tri1a/Tri1b (Fig. 3a–f). This process may be reversible to a threshold level, ensuring that a sufficient number of tail fibre-binding interactions occurs to disrupt enough lateral dimers to break the triplex iris ring and begin the irreversible process of sheath contraction[32]. This would provide a 'checksum' mechanism for the baseplate and avoid premature triggering. It may also be used to ensure that the pyocin is positioned upright to prevent sideward, non-productive firing, as the arrangement of triggering tail fibres may also play a role in setting off the baseplate[32] (Fig. 4a). As the iris breaks, the triplexes may pull the four-helical bundles of attached ripcord proteins, lowering the final barrier to contraction (Fig. 4).

Indeed, changing the lateral dimer interface between PA0618 Tri1a and Tri1b subunits alters the firing characteristics of pyocin. Owing to the presence of a histidine residue in the interface, pyocins are sensitive to an acidic environment and their contraction is triggered by pH 3.4 (Fig. 3g), at which the histidine is protonated to disturb the dimer interface (Fig. 3f). We engineered a mutant, H257F, in which the histidine is replaced by a phenylalanine. We found that the mutant is more tolerant to acid than the WT pyocin, whereas cryo-EM images show that purified WT and H257F both contain a similar percentage

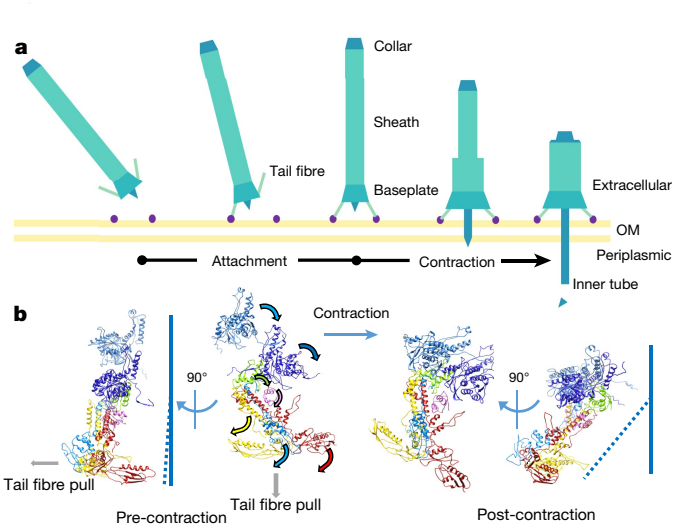


Fig. 4 | Baseplate transition from the pre-contracted to the post-contracted state. a, Illustration of a pyocin landing on a bacterial cell and firing. Release of the spike and hub following injection is postulated on the basis of the lack of these structures on contracted particles that we observed in vitro. OM, outer membrane. **b**, Ribbon diagram of the conserved baseplate components and sheath proteins in their pre-contracted and post-contracted states. Ripcord is believed to travel with the inner tube during the power stroke and therefore is not a conserved component of the baseplate after contraction. Arrows denote potential movements for subunits in the same colours, respectively.

of pre-contraction pyocins at neutral pH (64% and 67%, respectively); however, at pH 3.4, cryo-EM images show that far more WT than H257F pyocins have been triggered into the post-contraction state (9% and 36% of the remaining pre-contraction pyocins, respectively) (Fig. 3g). Two other PA0618 mutations at the lateral dimer interface, S250A and A254C (Fig. 3f), resulted in either defective assembly (S250A) or premature firing (A254C) of the particles.

Among the known structures of similar contractile systems, the sheath and sheath initiator are both conserved, which indicates that key aspects of the contraction mechanism are conserved[6,8]. Whether these similarities also extend to T6SS baseplate proteins is not known in the absence of atomic structures. Thus far, however, the four-helical bundle motif of the ripcord protein described here seems to be unique to pyocin R2 and related phages (for example, PS17 phage) based on genetic information. Indeed, a search in other contractile tail-like systems did not yield any four-helical bundle orthologues, although all contain orthologues for the triplex proteins. This unique triggering system puts pyocin R2 in a special position among contractile structures, perhaps due to its minimal nature that precludes sophisticated and sizeable insertion domains like those of T4.

R-type pyocins and related R-type bacteriocins are being developed as a new class of antimicrobials[2–4,28,29]. A unique feature of these structures is that highly specific target recognition conferred by receptor-binding proteins is directly coupled to the mechanism of action. This exquisite specificity allows selective killing of pathogens without the unintended consequences of off-target effects such as dysbiosis[33], and without selecting for transmissible antibiotic resistance in off-target bacterial species or strains. Engineering receptor-binding proteins to alter R-type pyocin-binding specificity has already been demonstrated[3,4,34]. We now have the information required to fine-tune the triggering mechanism through structure-guided alterations at key interaction points between baseplate components. For applications that require precise ablation of pathobionts from complex bacterial ecosystems, a less sensitive trigger would minimize off-target effects and would only set off the killing mechanism when tightly bound to

the correct bacterial cell. Conversely, in full-blown infections such as septicaemia, in which a single pathogen has grown to high density, a more sensitive trigger would allow for more efficient killing on collision with the target. The adaptability of contractile injection systems and the modularity of receptor-binding proteins, both honed over eons of evolution, provide an opportunity to engineer precision antibiotics for human and animal health.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2186-z.

1. Ge, P. et al. Atomic structures of a bactericidal contractile nanotube in its pre- and postcontraction states. *Nat. Struct. Mol. Biol.* **22**, 377–382 (2015).
2. Scholl, D. Phage tail-like bacteriocins. *Annu. Rev. Virol.* **4**, 453–467 (2017).
3. Scholl, D. et al. An engineered R-type pyocin is a highly specific and sensitive bactericidal agent for the food-borne pathogen *Escherichia coli* O157:H7. *Antimicrob. Agents Chemother.* **53**, 3074–3080 (2009).
4. Williams, S. R., Gebhart, D., Martin, D. W. & Scholl, D. Retargeting R-type pyocins to generate novel bactericidal protein complexes. *Appl. Environ. Microbiol.* **74**, 3868–3876 (2008).
5. Taylor, N. M. et al. Structure of the T4 baseplate and its function in triggering sheath contraction. *Nature* **533**, 346–352 (2016).
6. Leiman, P. G. et al. Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. *Proc. Natl Acad. Sci. USA* **106**, 4154–4159 (2009).
7. Kudryashev, M. et al. Structure of the type VI secretion system contractile sheath. *Cell* **160**, 952–962 (2015).
8. Basler, M., Pilhofer, M., Henderson, G. P., Jensen, G. J. & Mekalanos, J. J. Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature* **483**, 182–186 (2012).
9. Leiman, P. G. & Shneider, M. M. Contractile tail machines of bacteriophages. *Adv. Exp. Med. Biol.* **726**, 93–114 (2012).
10. Clemens, D. L., Ge, P., Lee, B. Y., Horwitz, M. A. & Zhou, Z. H. Atomic structure of T6SS reveals interlaced array essential to function. *Cell* **160**, 940–951 (2015).
11. Böck, D. et al. In situ architecture, function, and evolution of a contractile injection system. *Science* **357**, 713–717 (2017).
12. Ho, B. T., Dong, T. G. & Mekalanos, J. J. A view to a kill: the bacterial type VI secretion system. *Cell Host Microbe* **15**, 9–21 (2014).
13. Mougous, J. D. et al. A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science* **312**, 1526–1530 (2006).
14. Stover, C. K. et al. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**, 959–964 (2000).
15. Aksyuk, A. A. et al. The tail sheath structure of bacteriophage T4: a molecular machine for infecting bacteria. *EMBO J.* **28**, 821–829 (2009).
16. Kostyuchenko, V. A. et al. The tail structure of bacteriophage T4 and its mechanism of contraction. *Nat. Struct. Mol. Biol.* **12**, 810–813 (2005).
17. Leiman, P. G., Chipman, P. R., Kostyuchenko, V. A., Mesyanzhinov, V. V. & Rossmann, M. G. Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host. *Cell* **118**, 419–429 (2004).
18. Hu, B., Margolin, W., Molineux, I. J. & Liu, J. Structural remodeling of bacteriophage T4 and host membranes during infection initiation. *Proc. Natl Acad. Sci. USA* **112**, E4919–E4928 (2015).
19. Hu, B., Margolin, W., Molineux, I. J. & Liu, J. The bacteriophage t7 virion undergoes extensive structural remodeling during infection. *Science* **339**, 576–579 (2013).
20. Hatfull, G. F. Bacteriophage genomics. *Curr. Opin. Microbiol.* **11**, 447–453 (2008).
21. Hendrix, R. W., Hatfull, G. F. & Smith, M. C. Bacteriophages with tails: chasing their origins and evolution. *Res. Microbiol.* **154**, 253–257 (2003).
22. Chen, Z. et al. Cryo-EM structure of the bacteriophage T4 isometric head at 3.3-Å resolution and its relevance to the assembly of icosahedral viruses. *Proc. Natl Acad. Sci. USA* **114**, E8184–E8193 (2017).
23. Jiang, F. et al. Cryo-EM structure and assembly of an extracellular contractile injection system. *Cell* **177**, 370–383.e15 (2019).
24. Desfosses, A. et al. Atomic structures of an entire contractile injection system in both the extended and contracted states. *Nat. Microbiol.* **4**, 1885–1894 (2019).
25. Chang, Y. W., Rettberg, L. A., Ortega, D. R. & Jensen, G. J. In vivo structures of an intact type VI secretion system revealed by electron cryotomography. *EMBO Rep.* **18**, 1090–1099 (2017).
26. Hu, B. et al. Visualization of the type III secretion sorting platform of *Shigella flexneri*. *Proc. Natl Acad. Sci. USA* **112**, 1047–1052 (2015).
27. Low, H. H. et al. Structure of a type IV secretion system. *Nature* **508**, 550–553 (2014).
28. Kirk, J. A. et al. New class of precision antimicrobials redefines role of *Clostridium difficile* S-layer in virulence and viability. *Sci. Transl. Med.* **9**, eaah6813 (2017).
29. Ritchie, J. M. et al. An *Escherichia coli* O157-specific engineered pyocin prevents and ameliorates infection by *E. coli* O157:H7 in an animal model of diarrheal disease. *Antimicrob. Agents Chemother.* **55**, 5469–5474 (2011).
30. Browning, C., Shneider, M. M., Bowman, V. D., Schwarzer, D. & Leiman, P. G. Phage pierces the host cell membrane with the iron-loaded spike. *Structure* **20**, 326–339 (2012).
31. Kageyama, M., Ikeda, K. & Egami, F. Studies of a pyocin. Iii. Biological properties of the pyocin. *J. Biochem.* **55**, 59–64 (1964).
32. Crawford, J. T. & Goldberg, E. B. The function of tail fibers in triggering baseplate expansion of bacteriophage T4. *J. Mol. Biol.* **139**, 679–690 (1980).
33. Gebhart, D. et al. A modified R-type bacteriocin specifically targeting *Clostridium difficile* prevents colonization of mice without affecting gut microbiota diversity. *mBio* **6**, e02368-14 (2015).
34. Scholl, D., Gebhart, D., Williams, S. R., Bates, A. & Mandrell, R. Genome sequence of *E. coli* O104:H4 leads to rapid development of a targeted antimicrobial agent against this emerging pathogen. *PLoS One* **7**, e33637 (2012).

Methods

Purification of pyocins for cryo-EM

The pyocin sample was prepared from the *P. aeruginosa* strain PAO1 as previously described[1,4]. Briefly, a crude pyocin sample was first prepared from PAO1 and harvested by high-salt precipitation and differential centrifugation. This crude sample was further purified using a 10–50% sucrose (w/v) gradient at 77,000*g* for 1.5 h at 4 °C. After centrifugation, one band was visible at about the 25% position and was extracted gently by fractionation with a 100-µL pipette from the top of the centrifuge tube along its side. The extracted sample was then diluted to a final volume of 4 mL with Tris buffer (10 mM Tris and 130 mM NaCl, pH 7.4). The diluted sample was concentrated using a 100-kDa Amicon molecular filter to about 50 µL. This dilution concentration step was repeated three more times in the same filter as a means of dialysis, ending up with a final sample volume of 50 µL for cryo-EM imaging.

Cloning and purification of proteins for crystallography

Bioinformatic analysis with the help of HHpred[35] was used to identify the *P. aeruginosa* PAO1 genes encoding the central spike protein (PA0616) and Tri1 (PA0618) in the R2 pyocin cluster of *P. aeruginosa* PAO1. The proteins were amplified with the primers given in Supplementary Table 1. PA0616 was cloned into the pEEv3 vector (a pET23d derivative with a TEV protease cleavage site downstream from the N-terminal His-tag) using the BamHI and HindIII restriction sites. PA0618 was cloned into the standard pET23 vector using the NdeI and XhoI restriction sites to give rise to a protein with a non-cleavable C-terminal His-tag.

Both proteins were expressed in *Escherichia coli* B834 (DE3) cells grown in 2×TY medium at 37 °C and aerated at 200 rpm to an optical density at 600 nm (OD_{600}) of 0.6 and induced with the addition of 1 mM IPTG. The temperature of the culture was lowered to 18 °C and protein expression was carried out overnight. The cells were pelleted by centrifugation for 10 min at 8,000*g* at 4 °C and then lysed by sonication. The soluble fraction was separated by centrifugation for 15 min at 25,000*g* at 4 °C and loaded on the IMAC resin (GE Health Science). The IMAC affinity chromatography was performed in 10 mM Tris-Cl pH 8.0 with a 0–300 mM imidazole gradient. Fractions containing the protein were dialysed into 10 mM Tris-Cl pH 8.0 overnight. The TEV protease in a 1:10 mass-to-mass ratio (TEV protease:target protein) was added to the PA0616 sample, and proteolysis of the N-terminal His-tag was performed simultaneously with a dialysis into 10 mM Tris-Cl pH 8.0 overnight at room temperature. The dialysis of PA0618 was performed overnight at 4 °C. The subsequent steps of protein purification were the same for both proteins. The samples were loaded onto a MonoQ 10/100 GL column equilibrated with 10 mM Tris-Cl pH 8.0 and eluted with a 0–1 M NaCl gradient. Fractions containing the proteins of interest were then subject to size-exclusion chromatography on a Superdex 200 10/300 GL column in the buffer containing 10 mM Tris-Cl pH 8.0 and 150 mM NaCl. Fractions containing the proteins were pooled together and concentrated while the buffer was changed to 20 mM Tris-Cl pH 8.0.

Crystallization and structure determination of PA0616 and PA616d

For crystallization, PA0616 is concentrated to 18 mg/mL and subjected to a sparse matrix random screen using crystallization kits produced by Jena BioSciences. Best crystals of PA0616 were obtained in 60% PEG-400, 100 mM Na_2SO_4, 100 mM Bis-Tris buffer at pH 8.5 in hanging drop. Several data sets were collected at the beamlines PX-I and PX-III of the Swiss Light Source (SLS) and the best data set was used in structure refinement (Extended Data Table 2). The crystals did not require a special cryoprotectant and could be flash frozen directly from the drop.

The structure of PA0616 was solved by molecular replacement using the P2 phage gpV[30] as a search model (the two proteins display 31% sequence identity). The PA0616 unit cell contained six trimers in the asymmetric unit. The gpV search model contained the OB-fold and the first three strands of the β-helix. The initial solution was found with the help of Molrep[36]. It was rigid body refined with Refmac5[37]. Ten cycles of density modification by solvent flattening and non-crystallographic averaging with Parrot[38] improved the density to a point where the OB-fold and the first three β-strands could be interpreted in terms of PA0616 amino acids. A new model of 1 chain was then superimposed onto the other 17 chains comprising the asymmetric unit. A new round of rigid body refinement was performed with Refmac5[37]. The new density was subjected to an additional 15 cycles of solvent flattening and non-crystallographic averaging with Parrot. The new map was interpretable throughout and the model could be built for all but the last five amino acids, comprising the Fe-binding site. This region of the model was disordered in all other data sets of PA0616 (including different space groups).

To reveal the structure of the Fe-binding site, we designed a shortened mutant of PA0616 (called PA0616d) that comprised 90 C-terminal residues (amino acids 96–185) of the full-length protein. PA0616d was PCR-amplified using primers given in Supplementary Table 1 and cloned into pEEv3 so that it carried a cleavable His-tag at its N terminus. Expression and purification were performed in a way similar to the full-length protein. PA0616d was then concentrated to 15 mg/mL and subjected to a sparse matrix random screen using crystallization kits produced by Jena BioSciences. Best crystals of PA0616d were obtained in 34% PEG-400, 200 mM Na_2SO_4, 100 mM Na-acetate buffer at pH 5.0 in hanging drop. The crystals did not require a special cryoprotectant and could be flash frozen directly from the drop. The crystallographic data were collected at the beamline PX-I of the SLS. The structure was solved by molecular replacement using the middle part of the overlapping fragment (residues 120–179) of the full-length PA0616 with the help of the Phaser program[39]. The N-terminal part of the PA0616d β-helix was substantially different to that of the full length. Instead of forming a compact structure, it was splayed out and opened like a flower. This structural difference made structure solution difficult, but thanks to the availability of very high-resolution data (Extended Data Table 2), after several rounds of model building and refinement coupled to density modification with Parrot, an interpretable density could be obtained for the rest of the protein (including the Fe-binding site).

The structures of PA0616 and PA0616d were refined with Coot[40], Refmac5[37] and Phenix[41]. The final models of PA0616 and PA0616d were deposited to the PDB under the accession numbers 4S37 and 4S36, respectively.

Production of the PA0618 fragment suitable for crystallographic analysis and the solution of the structure

Full-length PA0618 failed to crystallize. Thus, we subjected it to limited proteolysis by trypsin. Full-length PA0618 was digested with trypsin (trypsin:PA0618 ratio of 1:500) in high-salt buffer (20 mM Tris pH 8.0, 400 mM NaCl and 4 mM $CaCl_2$) for 75 min at room temperature. The reaction was quenched by the addition of PMSF to a final concentration of 2 mM. This procedure resulted in two stable fragments. We loaded the mixture onto a His-Trap column (GE Healthcare) and found that one of the fragments bound to the resin, suggesting that this fragment retained the His-tag and therefore constituted a C-terminal part of the protein (PA0618C). The fragment eluted from the column with a 10 mM Tris pH 8.0 buffer containing imidazole at 250 mM. The eluted protein was further purified by size-exclusion chromatography (Superdex 75 HiLoad 16/60), dialysed into 20 mM Tris pH 8.0 and 5 mM DTT overnight and then purified by anion-exchange chromatography (Mono Q 10/100). Fractions comprising the elution peak were pooled and purified by size-exclusion chromatography (Superdex 75 HiLoad 16/60, 10 mM Tris pH 8.0 and 150 mM NaCl) again. The purified PA0618 C terminus was concentrated to 40 mg/mL and screened for crystallization conditions using Jena BioSciences crystallization kits. Crystallization drops of the best crystallization condition carried a crust, which prevented optimal crystal growth. Lowering of the protein

concentration prevented the formation of the crust, but it also stopped crystal formation. The best crystals were obtained by streak seeding of drops containing PA0618C at 17 mg/mL and 17% PEG-5000 MME, 150 mM $NH_4CH_3CO_2$ and 100 mM MES pH 6.5.

Crystallographic data were collected at SLS beamlines PX-I and PX-III using well solution supplemented with 25% ethylene glycol as a cryo-protectant. The initial phases were obtained by a single-wavelength anomalous diffraction technique (SAD) using a $HoCl_2$ heavy-atom derivative. The crystals were soaked in the well solution containing $HoCl_2$ at 5 mM for 24 h. The model comprising residues 202–295 plus a His-tag was initially built by buccaneer[42] and completed manually. The structure of a better native data set was refined using Coot[40], Refmac5[37] and Phenix[41]. The structure is deposited to the PDB under the accession number 5CES. The resultant model of the C terminus fragment was then used in combination with the cryo-EM-derived structure to build the final model of PA0618.

Cryo-EM

An aliquot of 2.5 µL of the above purified pyocin sample was loaded onto a 'baked'[43] Quantifoil 1.2/1.3 m, 200 mesh grid, blotted for 4 s at force 1, then flash frozen with a Vitrobot Mark IV (FEI). Cryo-EM data were collected as movies in an FEI Titan Krios microscope (operated at 300 kV) equipped with a Gatan imaging filter (GIF) (the slit was not inserted) and a K2 Summit direct electron camera in counting mode using Leginon software package[44] for automation.

The target defocus value is set to 2.0 µm under focus. Each movie contains 50 frames with 5 frames per second with a total accumulated dosage of 60 electrons per $Å^2$. The dose rate is measured at 6 electrons per $Å^2$ per second in the Digital Micrograph software package.

Frames within each movie were aligned to correct for drift as previously described[45], except that an iterative alignment scheme, as previously described elsewhere[46], was used in addition to the original software. We output three averages of selected frames: 1st to 50th frames for particle selection, 3rd to 20th frames for refinement, and 3rd to 13th frames for final reconstruction.

The contrast transfer function (CTF) parameters of these movies were determined from the averages with the 1st to 50th frames by CTFFIND3[47]. The range for acceptable defocus values was set to be between 1 and 3 µm, whereas outliers were rejected.

Image processing and 3D reconstruction

As shown in Fig. 1b, particles at both pre-contraction and post-contraction states are present in the cryo-EM images and can be readily distinguished by eye. The two ends of each pyocin were manually selected as individual particles with EMAN[48] boxer and were kept together before 3D classifications (see below). The box size for these particles is 320 and 420 pixels for the pre-contraction and post-contraction states, respectively; their total numbers of particles are 43,934 and 36,116, respectively. The trunk portion of each pre-contracted pyocin was boxed with EMAN[48] helixboxer with a box width of 384 and was segmented according to a 10% overlapping scheme, and a total of 15,684 segments selected. Because both ends of the post-contraction state contained sufficient structural components from the trunk, we did not pursue a helical reconstruction for the trunk in the post-contraction state. Throughout these particle boxing processes, we only boxed particles that were not overlapping, not broken and not ice-contaminated.

Further image processing steps were performed with Relion v1.2[49]. The boxed particles were first subjected to 2D classification to eliminate poor particles and then to a 3D classification to separate the collar and baseplate for each state (6 classes for the pre-contraction state and 12 classes for the post-contraction state). After separation of the collar and the baseplate, for each end of a pyocin particle and for each state, a further 3D classification was done to again eliminate poor particles (five classes for the pre-contraction baseplate and eight classes for the rest).

The 3D models for the classes after this classification appeared to be mutually shifted along the particle axis. Therefore, the 3D model for all good classes is aligned to the one in the middle. The resulting shifts were projected to the 2D space of the original particles and were applied to the particle centres. The particles of the good classes were then extracted again using the translated coordinates.

The finally selected and re-extracted particles for each of the four structures were subjected to an auto-refinement in Relion v1.2. A sixfold rotational symmetry was applied to the reconstructions. The number of finally included particles is as follows: the pre-contraction collar: 4,109; the pre-contraction baseplate: 26,104; the post-contraction collar: 9,934; and the post-contraction baseplate: 15,582. The overall averaged resolutions of these final structures are estimated by ResMap[50] to be: the pre-contraction collar: 3.9 Å; the pre-contraction baseplate: 3.4 Å; and the post-contraction collar and baseplate: 3.5 Å (Extended Data Fig. 1).

Protein subunit identification and atomic modelling

A purified pyocin preparation was run 5 mm into a 10% SDS gel and stained with Coomassie blue. Each band containing a pyocin protein was excised, digested in-gel with trypsin and subjected to liquid chromatography–tandem mass spectrometry for sequencing analysis (conducted at the University of California Davis Proteomics Core). Proteins were then identified by comparing the mass spectrometry sequence fragments to the *P. aeruginosa* PAO1 sequence (Fig. 1e and Extended Data Table 3).

Atomic models were built ab initio with Coot[40]. We had to assign each of the above-identified protein candidates to a specific region in our cryo-EM density map. At 3.4 Å resolution, this is possible because our map has sufficient resolving power for chain tracing and identification of side chains (Fig. 1c and Extended Data Fig. 2). For each peptide chain, we meticulously compared its secondary structure to the secondary structure predictions of all candidate proteins and found the best match. Once the match was found, the sequence of the candidate protein was compared to the amino acid side chain features in the density map to register the sequence. Then the α-carbon positions were manually traced with the above identified sequence in mind for each of the 11 unique atomic models. The α-carbon trace was converted to a poly-alanine strand, and finally mutated to the correct amino acid sequence. Each of the side chains were manually inspected and fitted into the density along with additional attention to secondary structure conformations. Protein subunits were assembled into an asymmetric unit. Only protein subunits in one asymmetric unit were built. Symmetry-related copies were generated using the 'sym' command in UCSF Chimera[51] with C6 rotational symmetry.

Real-space model refinement

We carried out model refinement with the phenix.real_space_refine command of the Phenix package[41] using default settings in three steps. First, we refined each monomer model with the corresponding density cryo-EM map individually. Second, we combined all of the monomers in an asymmetric unit and refined it with the whole cryo-EM map to separate clashing atoms between adjacent monomers within the asymmetric unit. Third, the refined model of the asymmetric unit was used to generate the full model of either the baseplate or the collar using the 'sym' command in UCSF Chimera with C6 rotational symmetry. This full model was refined globally with the NCS restraints enabled to separate clashing atoms among asymmetric units. At each of the refinement steps, we manually inspected the models to assess the quality of the refinement, made manual adjustments and repeated the refinement steps until a final structure was reached. The refined models of the individual monomers and the full baseplate and collar were validated exhaustively with EMRinger score[52], Ramachandran plot, C-beta, map CC and MolProbity[53] and the results are tabulated in Extended Data Fig. 1.

Structure-based mutagenesis for PA0618

Mutations in the pyocin gene cluster were made in the *P. aeruginosa* strain PAO1 by two-step allelic exchange following the method previously described[4]. Briefly, ~1,000-bp regions of DNA containing mutations were amplified from PAO1 genomic DNA using overlapping PCR and cloned into pEX18Gm (gift from H. Schweizer) that was digested with KpnI and EcoRI. The constructs were transformed by electroporation into PAO1, which was plated on 50 μg/mL gentamicin to select for single crossover integrants. Gentamicin-resistant isolates were then picked, grown for 3 h in LB and plated on LB-sucrose plates to select for second crossover events. Colonies were picked and screened for gentamicin sensitivity. These were then sequence verified for correct mutations. The primers used for these constructs are shown in Supplementary Table 2.

Crude WT and H257F-mutant pyocin were prepared as described above. Both were diluted 1:10 in either Tris buffer (10 mM Tris and 130 mM NaCl, pH 7.4) or phosphate-citrate buffer (28 mM Na_2HPO_4 and 36 mM citric acid, pH 3.4) to adjust the pH. Cryo-EM imaging was done for each of the four resulting samples as described above. The numbers of pre-contraction and post-contraction pyocins were counted through visual inspection of about 200 images for each condition and their percentages were tallied in the bar graph in Fig. 3f.

PA0626-mutant construction

The PA0626 mutants 626TEV and 626ΔWL were made using the modified allelic exchange approach[54]. In brief, the whole WT R2 pyocin gene cluster resided on a pETcoco-1-based plasmid, pSW192. In each case, two overlapping ~1-kb-long fragments carrying the mutation were amplified using pSW192 as a template by the primers listed in Supplementary Table 2.

The donor vectors for the exchange pAK6 and pAK20 were assembled of pairs of fragments by NEBuilder reaction (New England Biolabs) on the backbone of pWM91 plasmid[55] in which the ampicillin resistance gene was replaced with the kanamycin resistance gene. The recipient plasmid pSW192 was maintained in the $RecA^+$ *E. coli* 4 s strain[56]. The donor vectors were transformed into the MFDpyr *E. coli* strain[57], and conjugation between the donor and the acceptor strains was performed on LB agar plates overnight at 37 °C. Selection for the recombination products pSW192:pAK6 and pSW192:pAK20 was done on LB agar plates supplemented with kanamycin at 50 μg/mL. Counterselection for excision products, pSW192 and a mutation-carrying plasmid, was done on agar plates with 1% tryptone, 0.5% yeast extract and 5% sucrose (MilliporeSigma) overnight at room temperature. Colony screening was done by PCR. Plasmid identities were confirmed by sequencing.

626TEV in vivo digestion

For co-expression of pyocins with TEV protease, pS626TEV and pBAD24-based plasmid pB^hTEV coding TEV protease under the control of the arabinose-inducible promoter were co-transformed into the *E. coli* BL21ΔAraΔFhuA strain alongside with all necessary control combinations (WT pyocin coding plasmid pSW192 and pB^hTEV, pSW192 and pBAD24, and pS626TEV and pBAD24). Clones were selected on agar plates for ampicillin and chloramphenicol resistance, grown in liquid LB medium supplemented with ampicillin at 100 μg/mL and chloramphenicol with 10 μg/mL at 37 °C, induced with arabinose (0.01% and 0.03%) and incubated overnight at 30 °C. Fresh lysates were cleared from debris for 5 min at 15,000*g* in a microcentrifuge, tested for killing activity on *P. aeruginosa* 13 s strain lawns[58] in a spot assay and visualized on a JEM-2100 electron microscope (JEOL) after staining with 2% uranyl acetate.

Recombinant pyocin production

One litre of *E. coli* BL21ΔAraΔFhuA freshly transformed with each plasmid carrying WT or mutant pyocin gene clusters was grown in Lennox LB medium (Invitrogen) supplemented with chloramphenicol (11 μg/mL) in a 4 L Erlenmeyer flask at 37 °C and 240 rpm to an OD_{600} of 1.0. To induce pyocin expression, 0.5 mL of 20% arabinose were added, the temperature was decreased down to 30 °C, and cells were incubated overnight. To remove debris and residual bacteria, the lysate was centrifuged at 15,000*g* for 30 min in a F9-6x-1000 rotor (Thermo Fisher Scientific). The lysate was supplemented with 4 mg of DNase I and 4 mg of RNase A (MilliporeSigma). 30 g of solid NaCl and 100 g of PEG-8000 (Fisher Scientific) were added and dissolved and the lysate was incubated at 4 °C overnight to ensure the full precipitation of pyocins. Pyocins were pelleted at 15,000*g* for 15 min in a F9-6x-1000 rotor and pellets were resuspended in 10 mL of SM buffer (8 mM $MgCl_2$, 100 mM NaCl and 50 mM Tris-HCl pH 7.5) with DNase I and RNase A at 1 μg/mL each. The sample was extracted with 10 mL of chloroform, centrifuged in 50-mL falcon tubes at 15,000*g* for 15 min. The pyocin-containing aqueous phase was collected and pyocins were pelleted at 100,000*g* for 1 h in a Type 70 Ti rotor (Beckman Coulter). The pellet was dissolved in 0.5 mL of SM buffer with shaking at 100 rpm overnight at 4 °C. The sample was cleared from non-dissolvable material via centrifugation at 15,000*g* for 5 min in a microcentrifuge at room temperature and loaded on a step gradient of 10%, 20%, 30%, 40% and 50% sucrose (0.9 mL each) and centrifuged at 100,000*g* for 1 h in a SW 55 Ti rotor (Beckman Coulter). The upper pyocin-containing band was isolated and dialysed against two changes of 0.1× SM buffer. The concentration of the dialysed sample was determined on the basis of the adsorption at 280 nm and brought to 0.1 mg/mL. To get a control sample of fully contracted pyocins, 20 μL of 3 M glycine-HCl pH 2.5 were added to 4 mL of normalized pyocin sample, and contracted and non-contracted samples were dialysed against 10 mM NaCl and 10 mM phosphate buffer pH 7.0. The whole procedure was repeated several times to get at least three independently prepared samples for each circular dichroism experiment performed.

Circular dichroism

Pyocin contraction rates were measured on a JASCO J-815 CD spectrometer at 203 nm over a 67–74 °C range with an increment of 1 °C. The pyocin concentration for contracted and non-contracted samples was maintained equal to 0.1 mg/mL in 10 mM NaCl and 10 mM phosphate buffer pH 7.0. Each measurement took 25 min. The change of ellipticity for non-contracted pyocins was temperature dependent; however, it was not temperature dependent for contracted control samples. For data analysis, we averaged curves for each contracted sample and a resulting control curve subtracted from each contraction rate measurement of a cognate non-contracted sample. The first 2 min of every measurement were trimmed as they correspond to a sample being heated up to the desirable temperature and do not show pyocin contraction. Exponents were fitted into subtracted and trimmed data and the rate constants were determined with the help of MATLAB CFTool (MathWorks). Arrhenius modelling of activation energy was also done in MATLAB.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Cryo-EM maps and the associated atomic models have been deposited to the Electron Microscopy Data Bank (EMDB) and the PDB under the accession numbers EMD-20526/PDB: 6PYT (pre-contraction helical trunk), EMD-20643/PDB: 6U5B (pre-contraction baseplate), EMD-20646/PDB: 6U5H (pre-contraction hub in C3 symmetry), EMD-20644/PDB: 6U5F (pre-contraction collar), EMD-20647/PDB: 6U5J (post-contraction collar) and EMD-20648/PDB: 6U5K (post-contraction

baseplate), respectively. X-ray crystal structures have been deposited to the PDB under the accession numbers 5CES (PA0618 C-terminal domain), 4S36 (PA0616 C-terminal domain) and 4S37 (full-length PA0616). All other data are available from the corresponding authors on reasonable request.

Code availability

The modified version of MotionCorr 1 is available on GitHub, licensed under GPLv3 (gepeng1983/motioncorr1exp). Relion 1.2 with helical reconstruction patch is available on GitHub, licensed under GPLv2 (gepeng1983/relion12exp). A later version of Relion (1.4) with the same patch is also available on GitHub (gepeng1983/relion14exp).

35. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*. **33**, W244–W248 (2005).
36. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr*. **66**, 22–25 (2010).
37. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr*. **67**, 355–367 (2011).
38. Zhang, K. Y., Cowtan, K. & Main, P. Combining constraints for electron-density modification. *Methods Enzymol*. **277**, 53–64 (1997).
39. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr*. **40**, 658–674 (2007).
40. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr*. **66**, 486–501 (2010).
41. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr*. **66**, 213–221 (2010).
42. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr*. **62**, 1002–1011 (2006).
43. Miyazawa, A., Fujiyoshi, Y. & Unwin, N. Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423**, 949–955 (2003).
44. Suloway, C. et al. Automated molecular microscopy: the new Leginon system. *J. Struct. Biol*. **151**, 41–60 (2005).
45. Li, X. et al. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).
46. Banerjee, S. et al. 2.3 Å resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science* **351**, 871–875 (2016).
47. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol*. **142**, 334–347 (2003).
48. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol*. **128**, 82–97 (1999).
49. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol*. **180**, 519–530 (2012).
50. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
51. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem*. **25**, 1605–1612 (2004).
52. Barad, B. A. et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
53. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr*. **66**, 12–21 (2010).
54. Blomfield, I. C., Vaughn, V., Rest, R. F. & Eisenstein, B. I. Allelic exchange in *Escherichia coli* using the *Bacillus subtilis* sacB gene and a temperature-sensitive pSC101 replicon. *Mol. Microbiol*. **5**, 1447–1457 (1991).
55. Metcalf, W. W. et al. Conditionally replicative and conjugative plasmids carrying lacZ alpha for cloning, mutagenesis, and allele replacement in bacteria. *Plasmid* **35**, 1–13 (1996).
56. Prokhorov, N. S. et al. Function of bacteriophage G7C esterase tailspike in host cell adsorption. *Mol. Microbiol*. **105**, 385–398 (2017).
57. Ferrières, L. et al. Silent mischief: bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. *J. Bacteriol*. **192**, 6418–6427 (2010).
58. Scholl, D. & Martin, D. W. Jr. Antibacterial efficacy of R-type pyocins towards *Pseudomonas aeruginosa* in a murine peritonitis model. *Antimicrob. Agents Chemother*. **52**, 1647–1652 (2008).

Acknowledgements We thank X. Yu for advice in sample purification; UCLA students K. Wang, L. Nguyen, R. Chi, N. Poweleit and P. Graybeal and Beverly Hills High School students J. Gunn and L. Wang for picking particles; UCLA student E. Brown for video editing support; and D. Martin of AvidBiotics for discussion and support throughout this project. This research was supported in part by the NIH (R01GM071940 to Z.H.Z. and R21AI085318 to D.S.), the Swiss National Science Foundation (31003A_146284 to P.G.L.), and the Schaffer Family Foundation and Kavli Endowment (to J.F.M.). P.G. was supported in part by the American Heart Association Western States Affiliates Postdoc Fellowship (13POST17340020). We acknowledge the use of resources at the Electron Imaging Center for Nanomachines (EICN; supported by UCLA and by instrumentation grants from the NIH (1S10OD018111 and 1U24GM116792) and the NSF (DBI-1338135 and DMR-1548924)) and computation resource at the Extreme Science and Engineering Discovery Environment (XSEDE grant MCB140140 to Z.H.Z.). Recharge fees for access to the EICN facility for imaging the pyocin samples were partially defrayed by an award to Z.H.Z. from the UCLA CTSI core voucher program.

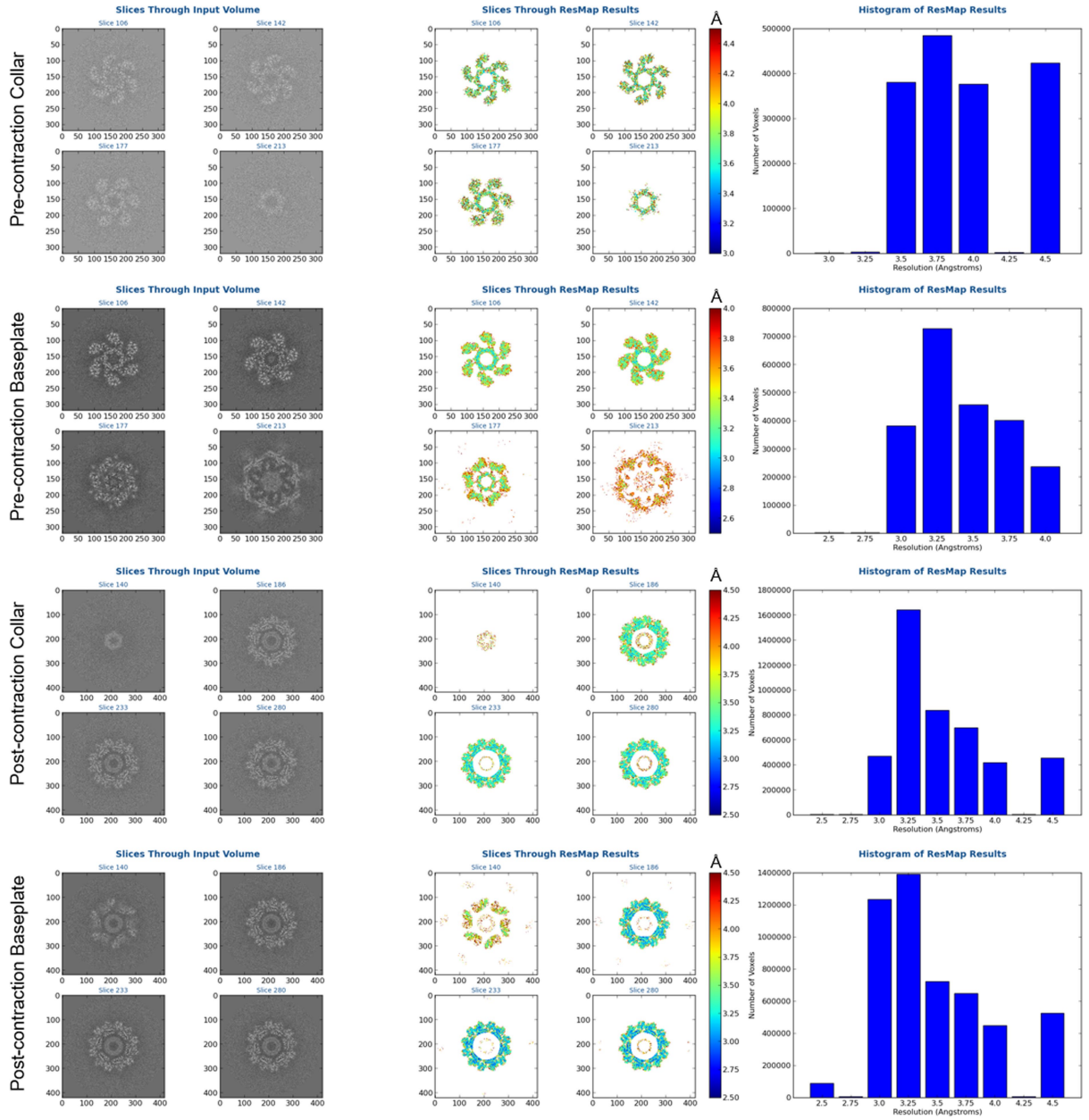
Author contributions Z.H.Z., J.F.M., P.G. and P.G.L. conceived the project. D.S., J.A. and K.D. prepared pyocin R2 samples used for high-resolution cryo-EM. P.G. and J.A. recorded the cryo-EM data. P.G. and Z.H.Z. processed the cryo-EM data. P.G., J.A. and P.G.L. built the atomic models using cryo-EM data. P.G., P.G.L., J.A. and Z.H.Z. analysed and interpreted the models. D.S., N.S.P. and U.C. created the pyocin mutants and examined their assembly properties and phenotypes. N.S.P. designed the functional assays and circular dichroism experiments to measure the activation energy of sheath contraction. M.M.S. created the expression constructs used for crystallography. C.B. and P.G.L. determined the crystal structure of the PA0616 spike protein. S.A.B. and M.P. determined the crystal structure of the C-terminal domain of PA0618. P.G., Z.H.Z., J.A., P.G.L., D.S. and J.F.M. wrote the paper. All authors contributed to the editing of the manuscript.

Competing interests J.F.M. is a cofounder, equity holder and a member of the Board of Directors of Pylum Biosciences, Inc., a biotherapeutics company in South San Francisco, CA, USA. D.S. is an equity holder of the same company.

Additional information

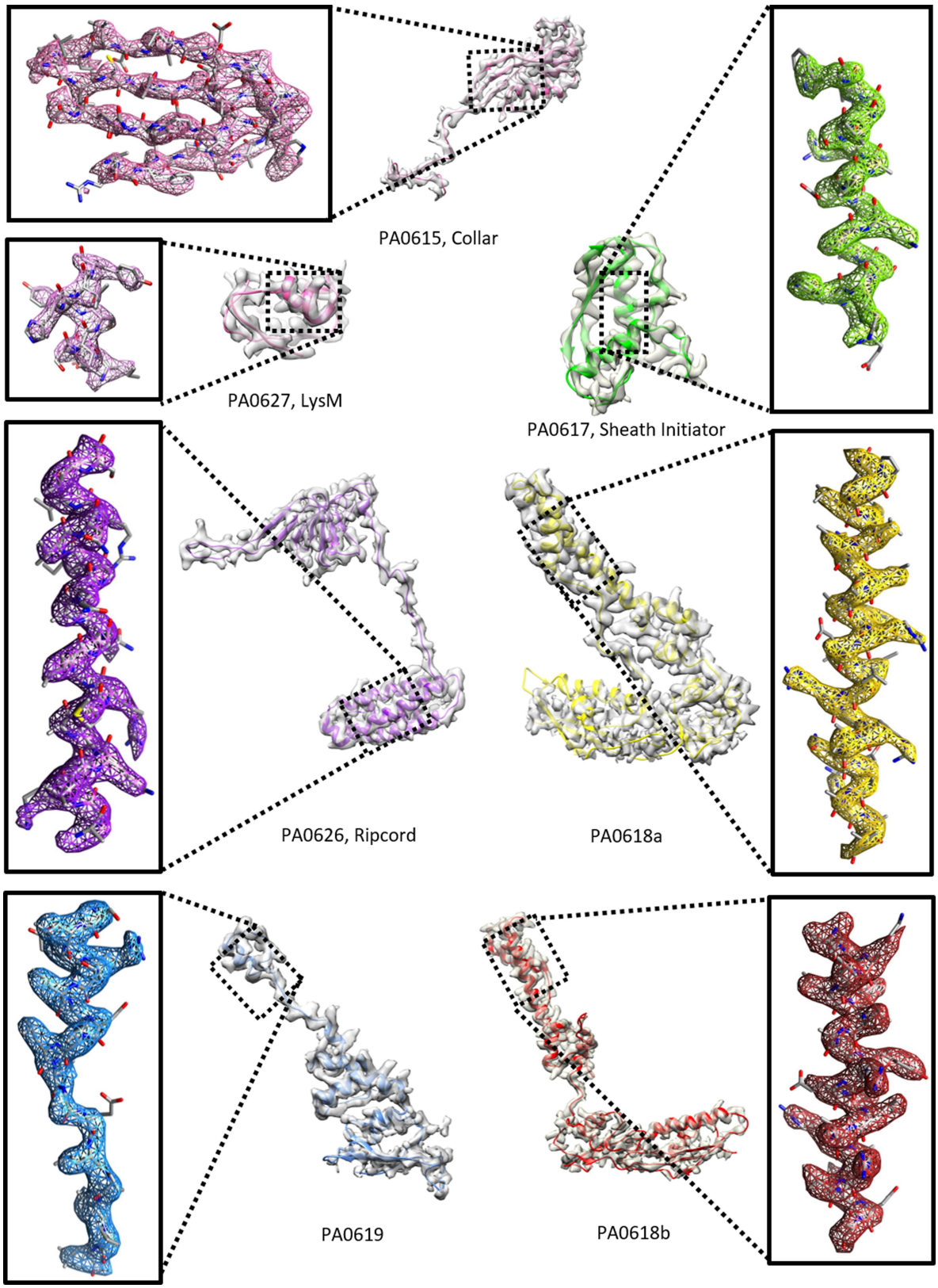
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2186-z.

Correspondence and requests for materials should be addressed to J.F.M. or Z.H.Z.

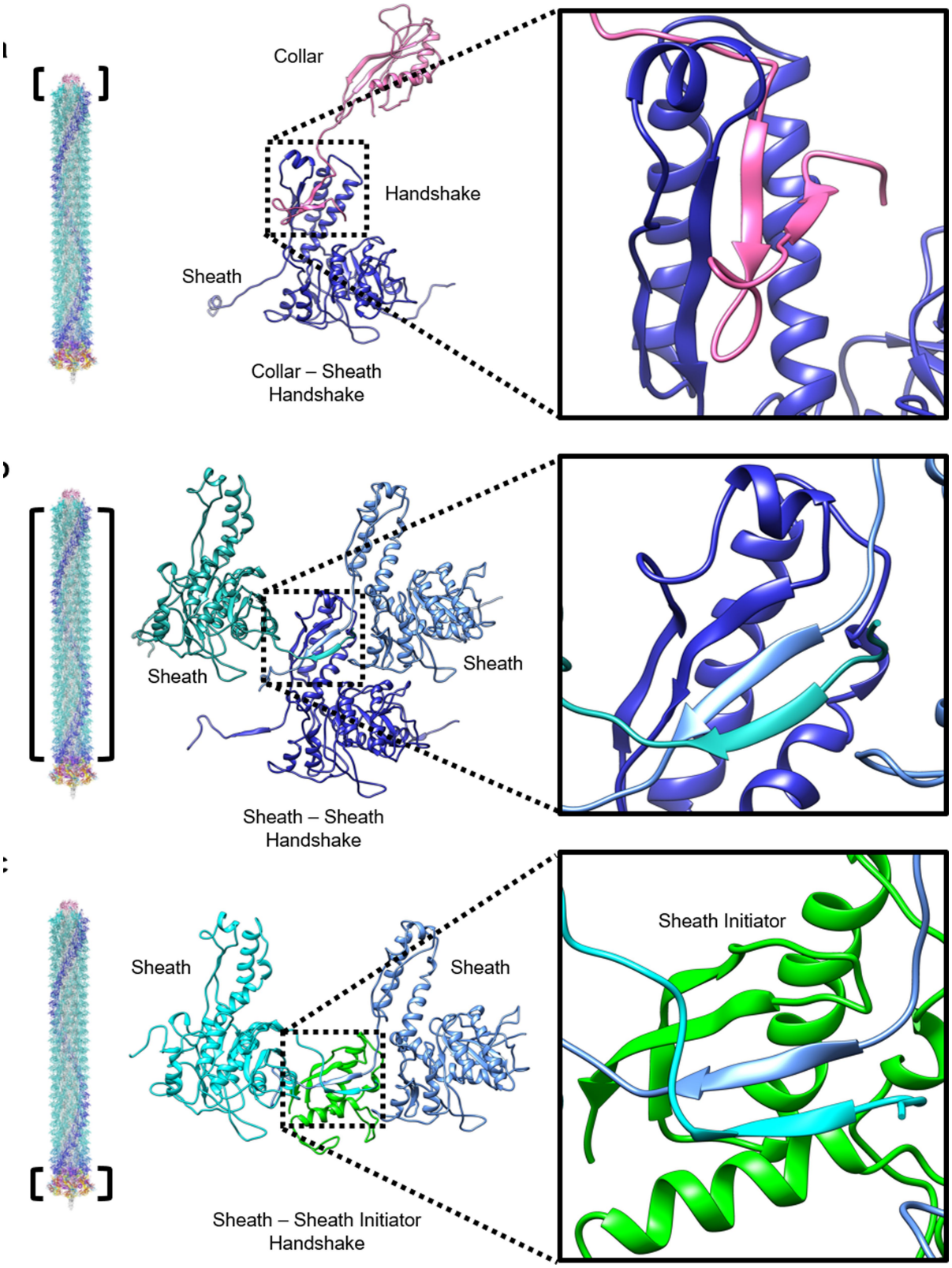


Protein		Ramachandran Outliers	Ramachandran Favored	Rotamer outliers	C-beta outliers	Bonds RMSD	Angles RMSD	Map CC	EMRinger Score
PA0615 / Collar		0%	90.9%	0.0%	0.0%	0.007	1.111	0.75	2.35
Baseplate Subunits	PA0617	0%	92.8%	0.0%	0.0%	0.006	0.925	0.82	2.62
	PA0618a	0%	91.7%	0.4%	0.0%	0.005	0.989	0.67	2.39
	PA0618b	0%	92.8%	0.4%	0.0%	0.005	0.920	0.73	2.01
	PA0619	0%	91.3%	0.0%	0.0%	0.005	1.040	0.71	2.54
	PA0626	0%	91.5%	0.0%	0.0%	0.006	0.965	0.82	3.26
	PA0627	0%	94.3%	0.0%	0.0%	0.005	0.957	0.81	4.59
Baseplate Average		0%	92.1%	0.2%	0.0%	0.005	0.967	0.77	2.35

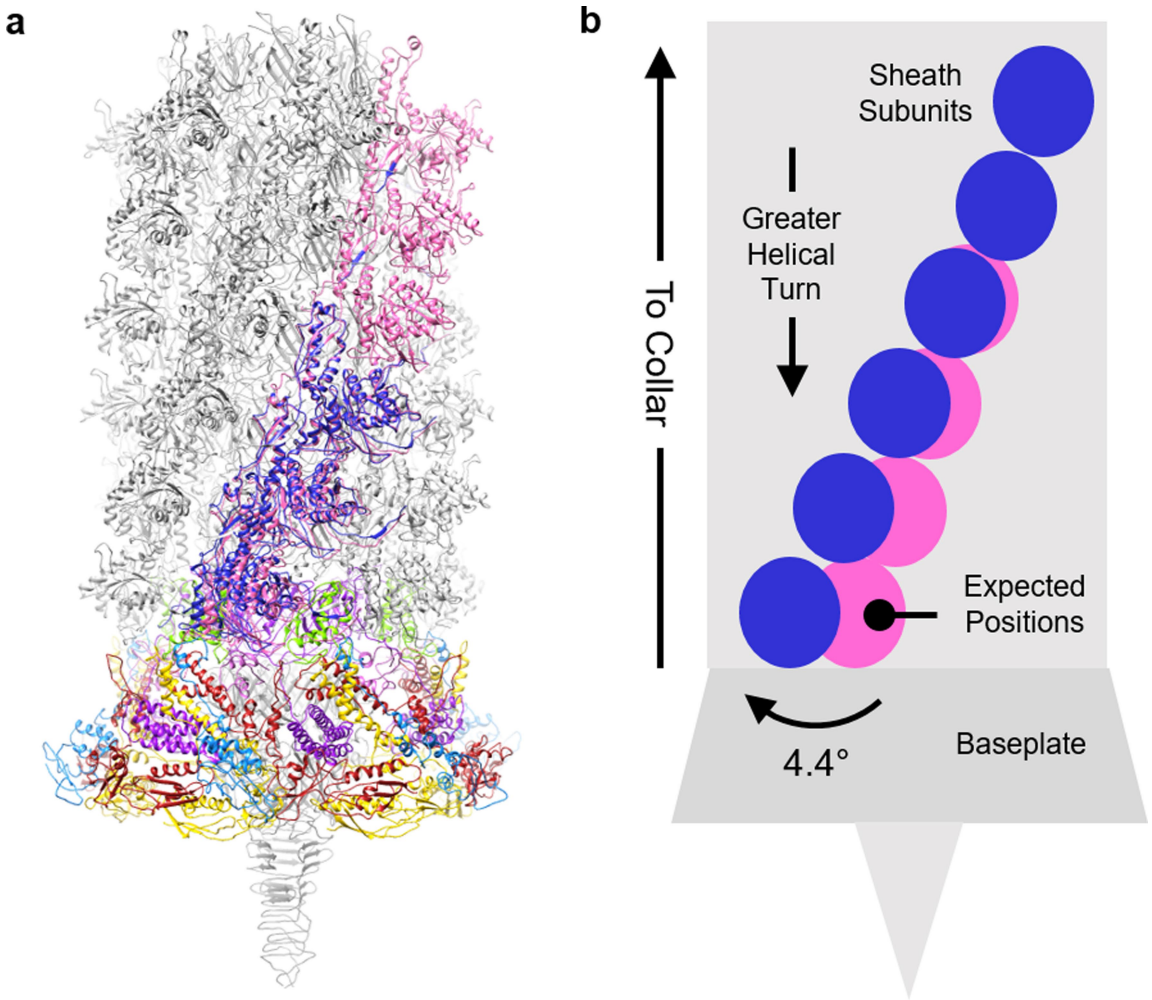
Extended Data Fig. 1 | Resolution assessment and model validation for the cryo-EM structures. ResMap results for the collar and baseplate regions of the pyocin reconstructions in pre-contraction and post-contraction states. Listed in the table are model validation statistics for the collar, baseplate subunits and average. For the 'Slices Through Input Volume' and the 'Slices Through ResMap Results' panels, the units on the axes are indexes for pixels.



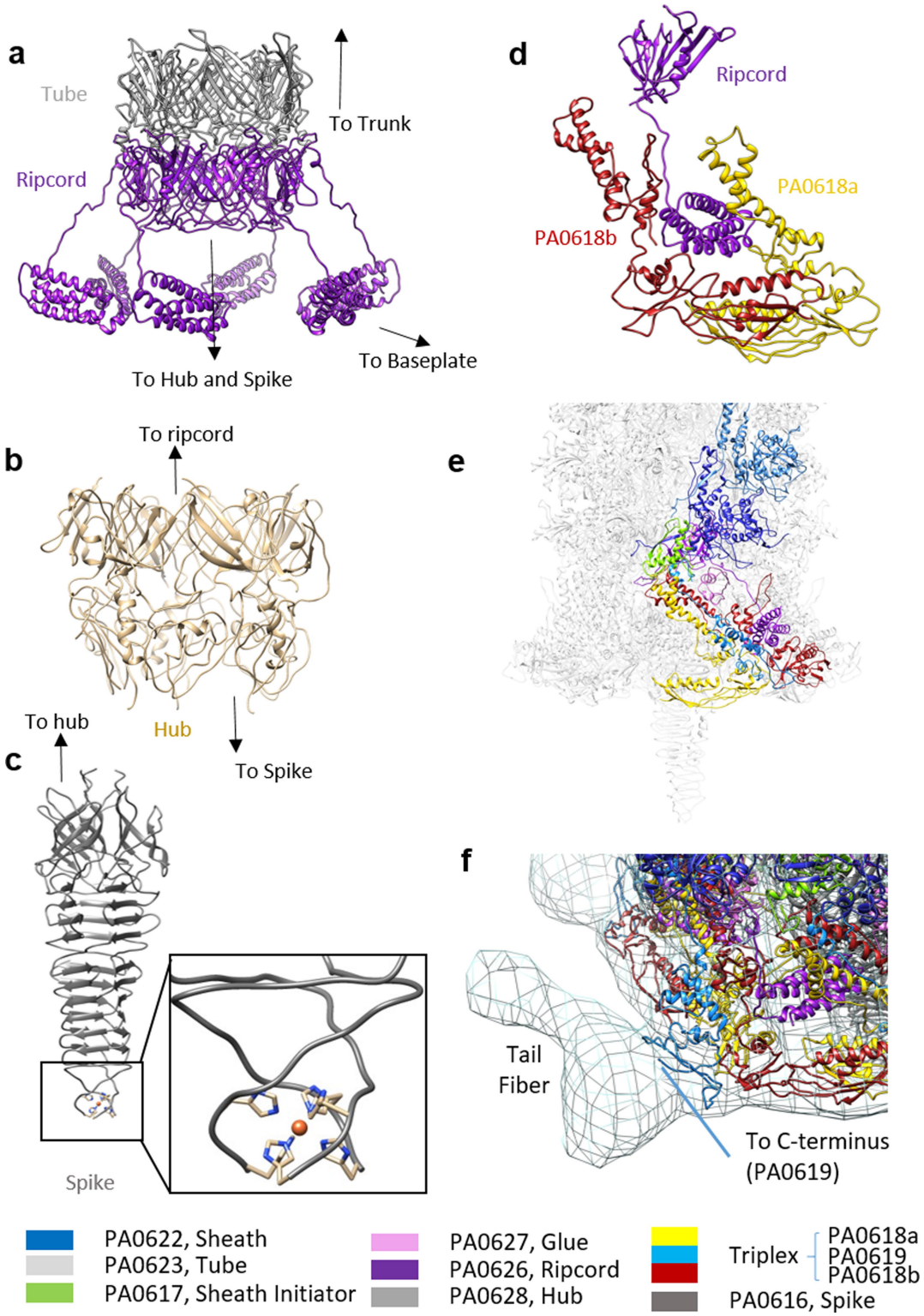
Extended Data Fig. 2 | Model assessment of pre-contraction pyocin subunits. For each of the collar and baseplate proteins, the cryo-EM density map is shown as semi-transparent grey superposed with its atomic model (ribbon). The close-up view of the box region shows the match of the density (wire frames) and the atomic model (sticks).



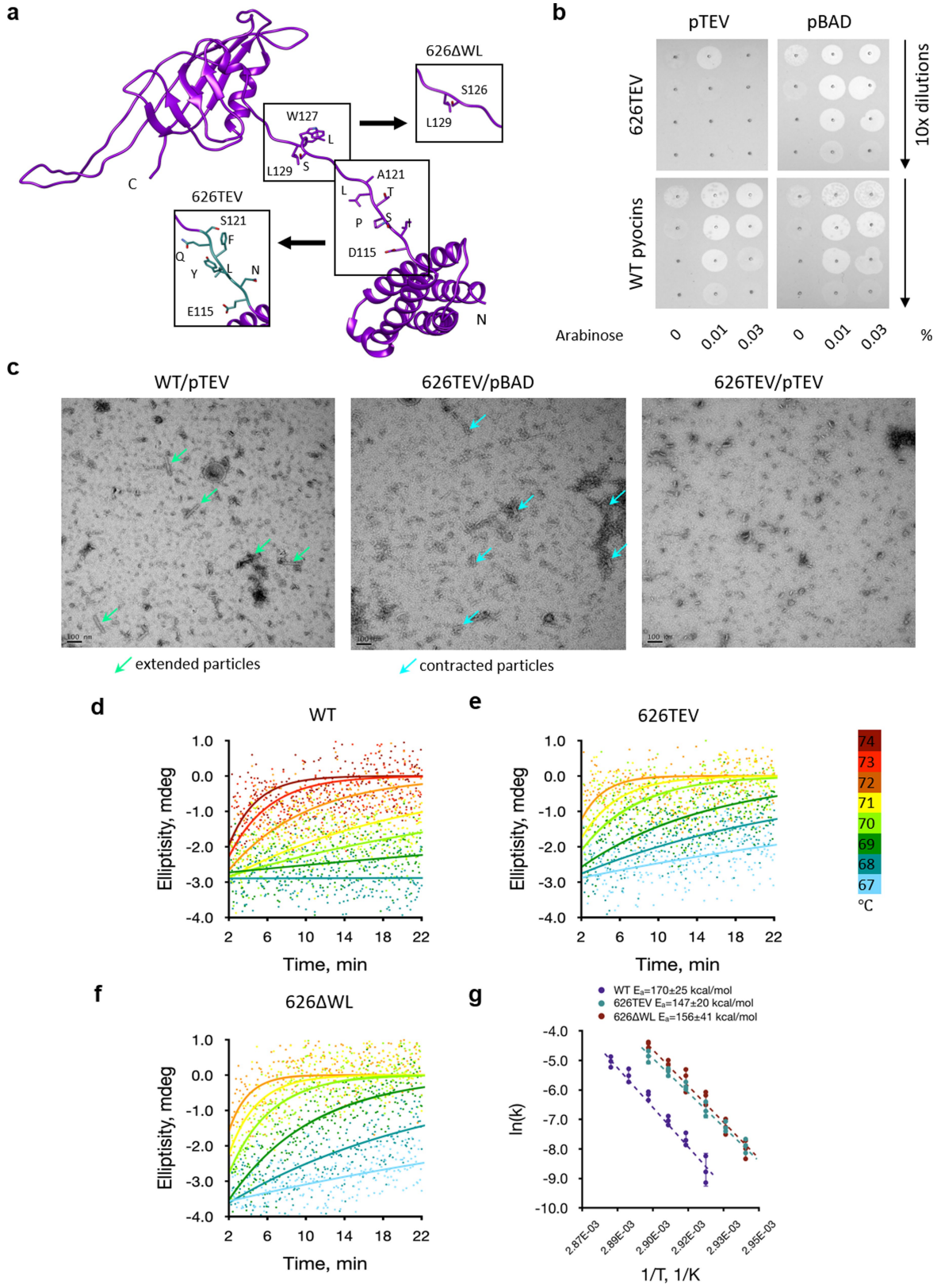
Extended Data Fig. 3 | Types of sheath handshakes in pyocin. Ribbon diagrams depicting the three types of handshake conformations in pyocin. **a**, Collar–sheath handshake. **b**, Sheath–sheath handshake. **c**, Sheath–sheath initiator handshake. All handshakes are composed of a four-stranded β-sheet.



Extended Data Fig. 4 | Trunk transitioning into the baseplate. a, Ribbon diagram depicting the lower portion of pyocin. The pink ribbons depict the expected positions of the sheath subunits according to the helical symmetry of the trunk. **b**, Schematic diagram depicting changes in the quaternary structure of the sheath subunits approaching the baseplate. The pink circles depict the expected positions of the sheath subunits according to helical symmetry of the trunk. The blue circles depict the actual positions with greater sequential helical turn, 4.4° at the last disc of the sheath.

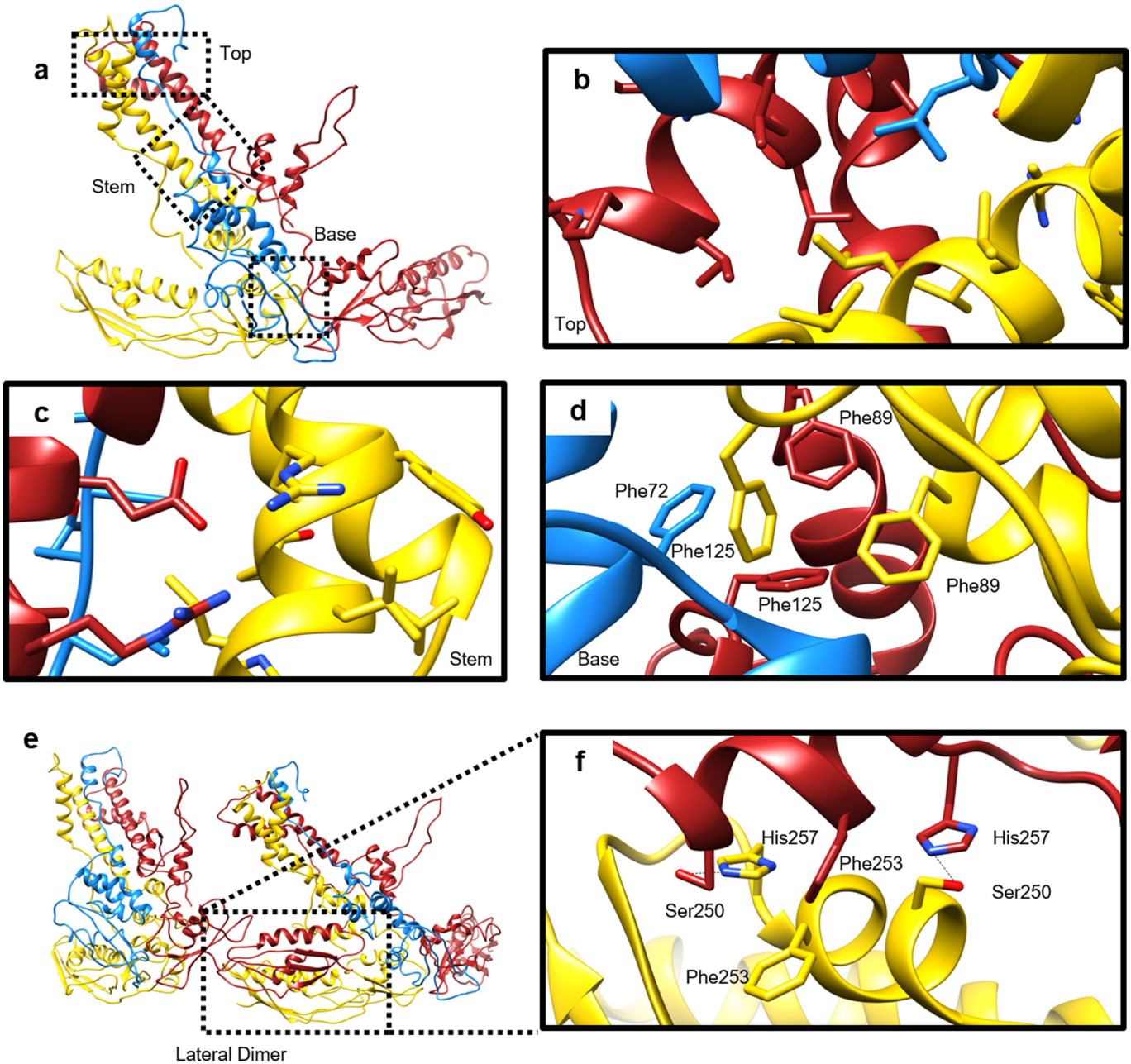


Extended Data Fig. 5 | Inspection of the baseplate. a, Ribbon diagram of the ripcord hexamer with the tube hexamer. **b**, **c**, Ribbon diagram of the hub (**b**) and the spike (**c**) (with the chelating site of its iron ion highlighted). **d**, Binding of the ripcord into triplexes. **e**, Ribbon diagram of the baseplate with one-sixth of its sixfold symmetric part highlighted in colours as in Fig. 1, showing the relative positions of each subunit. **f**, Baseplate ribbon model superimposed with a blurred cryo-EM density map of the proximal region of the tail fibre.

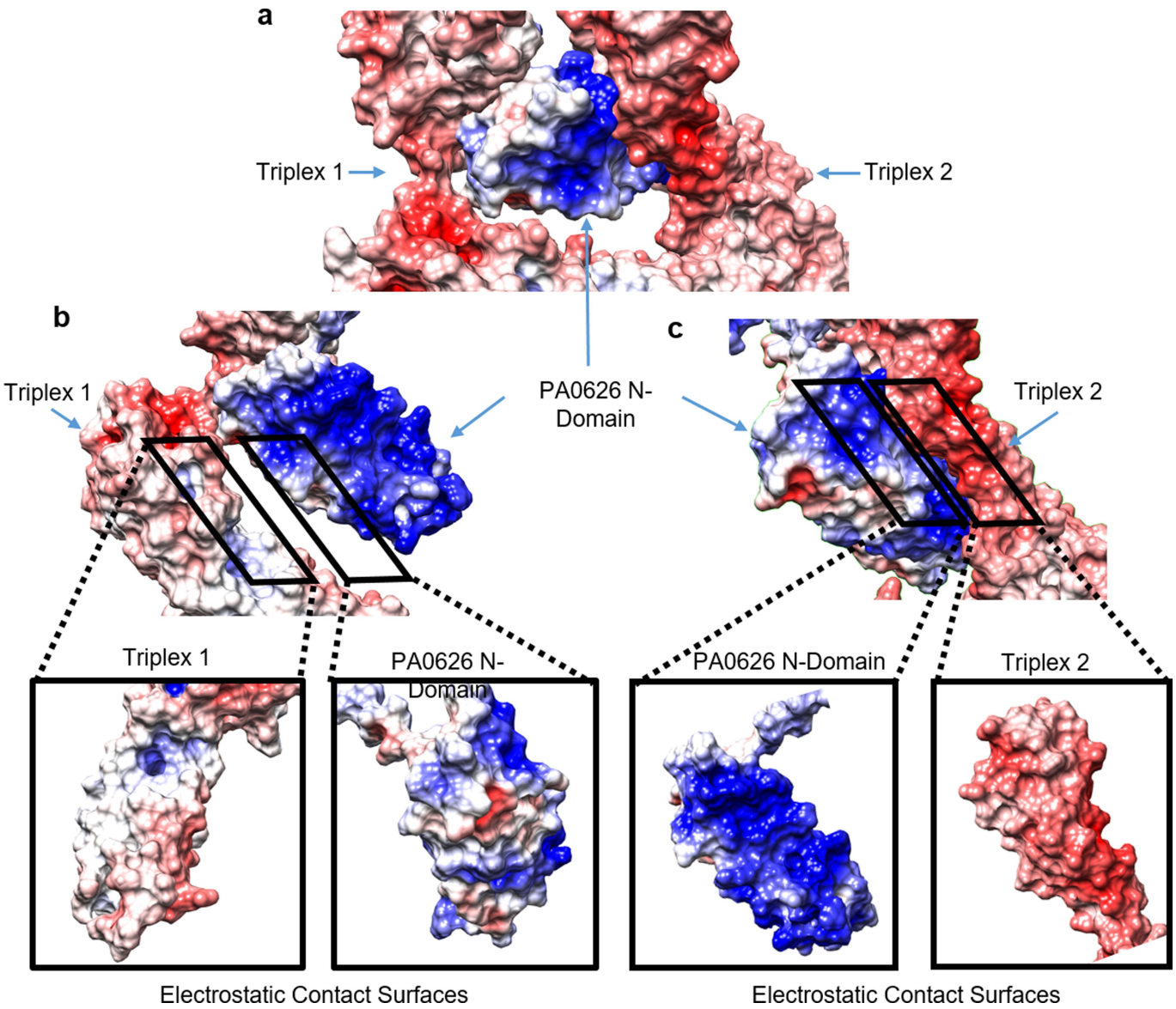


Extended Data Fig. 6 | See next page for caption.

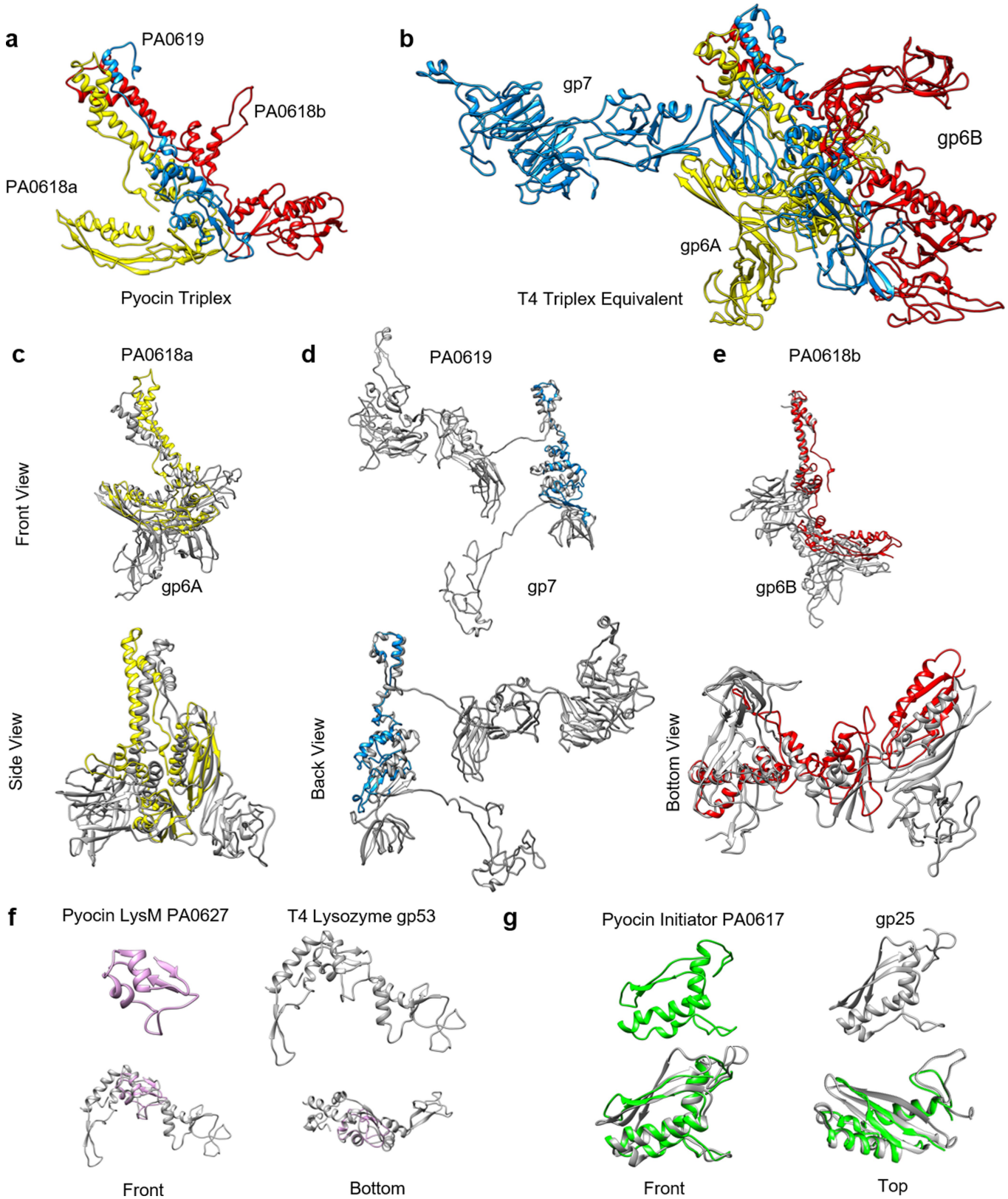
Extended Data Fig. 6 | Functional and morphogenetic implications of ripcord mutagenesis. a, Overview of ripcord mutagenesis. **b**, Co-expression of the WT pyocin and mutant 626TEV with the TEV protease. Pyocin killing activity in the lysates was assessed with the help of a spot assay with the *P. aeruginosa* 13 s strain as prey. Both pyocin and protease expression levels are arabinose dependent, with the rate of protease production being proportional to arabinose concentration and pyocin expression reaching the maximum at the lowest concentrations of arabinose tested (0.01%). Each experiment was repeated biologically three times (also for **c**–**g**). **c**, Representative negative staining EM images of the crude lysates shown in panel **b** induced with 0.01% arabinose. Despite showing killing activity in the lysates, no extended particles were found in the mutant 626TEV on EM grids. **d**–**f**, Temperature-dependent sheath contraction rates of the WT pyocins (**d**) and mutants (**e** and **f**) measured with the help of circular dichroism. **g**, The rate constants $k(T)$ of WT pyocins, 626ΔWL and 626TEV fitted to the Arrhenius model $k(T) = A \exp(-E_a/RT)$ where T is the absolute temperature, A is a temperature independent constant, E_a is the activation energy and R is the ideal gas constant. The dots on the graph are individual values for three biologically independent measurements of $\ln k(T)$, and the error bars show the 95% confidence interval calculated for them.



Extended Data Fig. 7 | Interactions important for triplex formation. **a**, Ribbon diagram of the atomic model of the pyocin triplex. **b**–**d**, The ribbon model with the depicted side chains in the top (**b**), stem (**c**) and base (**d**) regions of the triplex. In panel **d**, the phenylalanine pi-stacking coordination between PA0618a (yellow, Phe89 and Phe125), PA0618b (red, Phe89 and Phe125) and PA0619 (blue, Phe72) is shown. **e**, Ribbon model diagram of the lateral dimer. **f**, Close-up of the ribbon model diagram of the lateral dimer, highlighting the key interacting residues (Phe253–Phe253, His257–Ser250 and Ser250–His257).



Extended Data Fig. 8 | Electrostatic views of the ripcord handle. **a**, Electrostatic surface diagram of the ripcord with adjacent triplexes. **b**, **c**, Electrostatic properties of the interfaces between the ripcord and triplexes 1 and 2, respectively. The colour code corresponds to positive (blue), neutral (white) and negative (red).



Extended Data Fig. 9 | Comparison of related protein subunits from pyocin R2 and T4 phage. a, Ribbon diagram of the pyocin triplex. **b**, Ribbon diagram of the T4 triplex equivalent with subunits marked by corresponding colour to panel **a**. **c**, Ribbon diagram of pyocin PA0618a (yellow) and T4 gp6A (grey). **d**, Ribbon diagram of pyocin PA0619 (blue) and T4 gp7 (grey). **e**, Ribbon diagram of pyocin PA0618b (red) and T4 gp6B (grey). **f**, Ribbon diagram of pyocin PA0627 (pink) and T4 gp53 (grey). **g**, Ribbon diagram of pyocin PA0617 (green) and T4 gp25 (grey).

Extended Data Table 1 | Cryo-EM data statistics

	Pre-contraction Trunk (EMDB-20526) (PDB 6PYT)	Pre-contraction Baseplate (EMDB-20643) (PDB 6U5B)	Pre-contraction Hub (C3 sym.) (EMDB-20646) (PDB 6U5H)	Pre-contraction Collar (EMDB-20644) (PDB 6U5F)	Post-contraction Collar (EMDB-20647) (PDB 6U5J)	Post-contraction Baseplate (EMDB-20648) (PDB 6U5K)
Data collection and processing						
Magnification	130,000	130,000	130,000	130,000	130,000	130,000
Voltage (kV)	300	300	300	300	300	300
Electron exposure (e–/Å^2)	30	30	30	30	30	30
Defocus range (μm)	1.2-3.4	1.2-3.4	1.2-3.4	1.2-3.4	1.2-3.4	1.2-3.4
Pixel size (Å)	1.07	1.07	1.07	1.07	1.07	1.07
Symmetry imposed	C6 + helix	C6	C3	C6	C6	C6
Initial particle images (no.)	15,680	43,934	43,934	43,934	36,116	36,116
Final particle images (no.)	15,500	22,001	21,844	4,110	15,582	9,934
Map resolution (Å)	2.9	3.5	3.9	3.8	3.5	3.5
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.9-200	3.4-200	3.7-200	3.8-200	3.5-200	3.5-200
Refinement						
Initial model used (PDB code)	*de novo*	*de novo*	*de novo*	*de novo*	*de novo*	*de novo*
Model resolution (Å)	3.1	3.4	4.0	3.7	3.5	3.5
FSC threshold	0.5	0.5	0.5	0.5	0.5	0.5
Model resolution range (Å)	2.9-200	3.4-200	3.7-200	3.8-200	3.5-200	3.5-200
Map sharpening *B* factor (Å^2)	-80	-80	-80	-80	-80	-80
Model composition						
Non-hydrogen atoms	4,9920	103,392	7,326	107,550	59,634	110,142
Protein residues	6,624	13,608	954	14,232	7,878	14,520
Ligands	0	0	0	0	0	0
B factors (Å^2)						
Protein	45.39	41.54	55.81	13.62	37.18	76.64
Ligand						
R.m.s. deviations						
Bond lengths (Å)	0.009	0.005	0.005	0.007	0.005	0.005
Bond angles (°)	1.000	0.941	1.054	0.999	0.700	0.770
Validation						
MolProbity score	1.83	1.87	2.13	1.92	2.15	2.13
Clashscore	5.45	6.35	9.36	6.33	5.69	5.85
Poor rotamers (%)	0.94	0.17	0.39	0.38	3.03	3.00
Ramachandran plot						
Favored (%)	90.48	91.13	86.41	89.01	93.32	92.97
Allowed (%)	9.52	8.87	13.59	10.99	7.68	7.03
Disallowed (%)	0.00	0.00	0.00	0.00	0.00	0.00

FSC, Fourier shell correlation; R.m.s., root mean square.

Extended Data Table 2 | Crystallographic data statistics

Dataset name*	PA0618C $HoCl_2$	PA0618C	PA0616d	PA0616
Data collection				
Space group	$P4_3$	$P4_3$	$P6_322$	$P2_1$
Cell dimensions:				
a, b, c (Å)	102.67, 102.47, 45.76	72.61, 72.61, 46.52	46.68, 46.68, 144.55	94.56, 137.052, 164.44
α, β, γ (º)	90.00, 90.00, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 120.00	90.00, 106.88, 90.00
Wavelength (Å)	1.5352	0.91963	0.9763	1.6984
Resolution (Å)	50.0 – 2.80 (2.97 – 2.80)#	50.0 – 2.10 (2.23 – 2.10)	50.0 – 1.46 (1.55 – 1.46)	50.0 – 2.20 (2.33 – 2.20)
R_{meas} (%)	9.7 (58.3)	3.2 (73.6)	3.4 (53.4)	6.5 (95.2)
$CC_{1/2}$	99.9 (91.5)	100.0 (81.8)	100.0 (98.4)	99.8 (83.2)
I / σ_I	25.35 (4.65)	30.41 (2.61)	56.52 (5.35)	12.70 (1.33)
Completeness (%)	99.2 (96.3)	99.6 (97.8)	99.6 (97.7)	97.1 (92.3)
Redundancy	13.9 (13.5)	6.9 (6.7)	21.8 (16.0)	3.21 (2.95)
Anomalous signal$	1.770 (0.863)			
Refinement				
Software (version)		Phenix_refine (1.9_1692)	Phenix_refine (1.8.2_1309)	Phenix_refine (1.8.2_1309)
Resolution (Å)		50.0 – 2.10	50.0 – 1.46	50.0 – 2.20
No. unique reflections		27,538	16,615	203,160
No. atoms				
Protein		1,403	744	23,767
Ligand/ion		0	1	238
Water		58	118	2,432
Rwork / Rfree		0.190 / 0.236	0.149 / 0.188	0.163 / 0.208
B-factors:				
Protein (Å²)		74.92	28.64	54.02
Ligand/ion (Å²)		n/a	24.51	117.84
Water (Å²)		68.68	46.50	55.16
R.m.s. deviations				
Bond lengths (Å)		0.003	0.008	0.004
Bond angles (º)		0.705	1.274	0.889
Ramachandran plot:				
Favored (%)		97.77	98.86	96.03
Allowed (%)		2.23	0.0	3.47
Outliers (%)		0.0	1.14	0.50
PDB code		**5CES**	**4S36**	**4S37**

*A single crystal was used for each of the datasets.

#The statistics for the highest-resolution shell are given in parenthesis.

$As calculated by the program XDS (http://xds.mpimf-heidelberg.mpg.de/).

Extended Data Table 3 | Pyocin protein identification by mass spectrometry

MW (kDa)	Amino Acids	Gene Number (prf)	Gene Number (PA)	Copies (of monomer)
7.5	68	22	PA0627	Unknown
11.8	108	12	PA0617	6-18
18.1	168	18	PA0623	100-200
18.9	171	10	PA0615	Unknown
19.4	185	11	PA0616	6-18
20.0	177	14	PA0619	Unknown
31.3	290	21	PA0626	6-18
32.0	295	13	PA0618	6-18
35.9	329	23	PA0628	6-18
41.2	386	17	PA0622	100-200
77.7	745	20	PA0625	1

MW, molecular weight.

nature research

Corresponding author(s): Jeff F Miller, Z Hong Zhou

Last updated by author(s): Dec 16, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Leginon 3.3 (leginon.org);
Data analysis	Refmac5; Parrot; Phaser; Coot 0.8.9; Phenix 1.14-3260; CTFFind 3.1; MotionCorr 1 (with custom patch, available upon request); EMAN 1.81; Relion 1.2 (with custom IHRSR patch, available upon request, current version 1.4); Chimera 1.11.2; EMRinger (https://fraserlab.com/2015/02/18/EMringer/); MatLab (MathWorks)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

CryoEM maps and the associated atomic models have been deposited to EMDB, PDB under the accession numbers EMD-20526/PDB:6PYT (pre-contraction helical trunk), EMD-20643/PDB:6U5B (pre-contraction baseplate), EMD-20646/PDB:6U5H (pre-contraction hub in C3 symmetry), EMD-20644/PDB:6U5F (pre-contraction collar), EMD-20647/PDB:6U5J (post-contraction collar) and EMD-20648/PDB:6U5K (post-contraction baseplate), respectively. X-ray crystal structures have been deposited to PDB under the accession numbers 5CES (PA0618 C-terminal domain), 4S36 (PA0616 C-terminal domain) and 4S37 (full length PA0616). All other data are available from the corresponding authors upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The presented study is of structural biological nature. We averaged huge numbers of particles (See Ext Data Table 1) to reach the stated resolutions, which is then tested and confirmed in the work. For Figure 4f, the phenotypical evaluation of H257F mutant, 200 micrographs were taken for each condition and particles counted. For Ext Data Figure 6, samples were prepared independently three times, 10 micrographs were taken for each of them (b and c), giving us 0 pre-contraction and about 150 post-contraction pyocins for the WT and 0 pre- and about 150 post-contraction pyocins for the mutant analyzed in total. For (d-g), each experiment was repeated three times.
Data exclusions	No data were excluded.
Replication	Ext Data Fig 6: all experiments were repeated three times.
Randomization	The structural determination process is random by nature since the particles are randomly presented to us in solution.
Blinding	The structural determination process is blinded by nature since the particles are presented to us in solution without selection by prior knowledge.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

An open-source drug discovery platform enables ultra-large virtual screens

https://doi.org/10.1038/s41586-020-2117-z

Received: 5 March 2019

Accepted: 27 February 2020

Published online: 9 March 2020

Check for updates

Christoph Gorgulla[1,2,3]✉, Andras Boeszoermenyi[1,3,11], Zi-Fu Wang[1,11], Patrick D. Fischer[1,3,4], Paul W. Coote[1,3], Krishna M. Padmanabha Das[1,3], Yehor S. Malets[5,6], Dmytro S. Radchenko[5,6], Yurii S. Moroz[6,7], David A. Scott[1,3], Konstantin Fackeldey[8,9], Moritz Hoffmann[10], Iryna Iavniuk[5], Gerhard Wagner[1] & Haribabu Arthanari[1,3]✉

On average, an approved drug currently costs US$2–3 billion and takes more than 10 years to develop[1]. In part, this is due to expensive and time-consuming wet-laboratory experiments, poor initial hit compounds and the high attrition rates in the (pre-)clinical phases. Structure-based virtual screening has the potential to mitigate these problems. With structure-based virtual screening, the quality of the hits improves with the number of compounds screened[2]. However, despite the fact that large databases of compounds exist, the ability to carry out large-scale structure-based virtual screening on computer clusters in an accessible, efficient and flexible manner has remained difficult. Here we describe VirtualFlow, a highly automated and versatile open-source platform with perfect scaling behaviour that is able to prepare and efficiently screen ultra-large libraries of compounds. VirtualFlow is able to use a variety of the most powerful docking programs. Using VirtualFlow, we prepared one of the largest and freely available ready-to-dock ligand libraries, with more than 1.4 billion commercially available molecules. To demonstrate the power of VirtualFlow, we screened more than 1 billion compounds and identified a set of structurally diverse molecules that bind to KEAP1 with submicromolar affinity. One of the lead inhibitors (iKeap1) engages KEAP1 with nanomolar affinity (dissociation constant (K_d) = 114 nM) and disrupts the interaction between KEAP1 and the transcription factor NRF2. This illustrates the potential of VirtualFlow to access vast regions of the chemical space and identify molecules that bind with high affinity to target proteins.

Repeated optimization of lead compounds and late-stage failure of drug candidates are the primary causes of the longer development times and increased costs of drug development. Improving the quality of the initial lead compounds would minimize these lead optimization cycles and result in drug candidates that enter (pre-)clinical phases with greater specificity and higher affinity. Virtual screening to identify molecules that bind to a specified site on a receptor protein has become an important part of the drug discovery pipeline[2–5].

Current virtual screening paradigms routinely sample only a small fraction, on the order of 10^6–10^7 molecules, of the total chemical space of small organic compounds that are suitable for drug discovery, which have been estimated to encompass more than 10^{60} molecules[6].

However, the scale of a virtual screen is of central importance because the more compounds that are screened, the lower the rate of false positives and the more favourable the quality of the lead compounds (for example, molecules that bind with higher affinity). It was recently shown experimentally that ultra-large scale screening improves the rate of true positives[2]. Here we derived a probabilistic model of the true-positive rate as a function of the number of compounds screened; analysis of our ultra-large screen confirms that the docking score of the highest-scoring compounds improves with the scale. Increasing the scale of a virtual screen can improve the quality of initial hits in two distinct ways: by identifying hits with tighter binding affinity, which can result in lowered dosages and fewer off-target effects and by discovering compounds with more favourable pharmacokinetic and/or less inherent cytotoxic properties.

To increase the number of compounds evaluated in a virtual screen by orders of magnitude and make it accessible to any researcher, there is a need for a platform that can integrate all of the tasks in the virtual screening process. Such a platform should ideally scale linearly with the number of CPUs, efficiently handle billions of files, minimize input and output load, run robustly (for example, skip incorrectly encoded ligands, resist temporary input and output problems and resume after unexpected termination), run on any type of computing cluster

[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Harvard University, Boston, MA, USA. [2]Department of Physics, Faculty of Arts and Sciences, Harvard University, Cambridge, MA, USA. [3]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [4]Department of Pharmacy, Pharmaceutical and Medicinal Chemistry, Saarland University, Saarbrücken, Germany. [5]Enamine, Kyiv, Ukraine. [6]National Taras Shevchenko University of Kyiv, Kyiv, Ukraine. [7]Chemspace, Kyiv, Ukraine. [8]Zuse Institute Berlin, Berlin, Germany. [9]Institute of Mathematics, Technical University Berlin, Berlin, Germany. [10]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. [11]These authors contributed equally: Andras Boeszoermenyi, Zi-Fu Wang. ✉e-mail: cgorgulla@g.harvard.edu; hari@hms.harvard.edu

(including cloud platforms) and be user friendly and easy to use for non-computational scientists. Furthermore, to provide flexibility, a structure-based virtual screening platform should be able to interface with a variety of docking programs, support both rigid and flexible receptor docking, test multiple docking scenarios in a single workflow, allow for consensus and ensemble docking, and carry out multiple replicas of the same docking scenario. Lastly, to democratize access, facilitate widespread use and catalyse further development, such a platform would need to be open source.

With these requirements in mind, we designed VirtualFlow, an open-source platform that is able to screen chemical space on a large scale. Screening 1 billion compounds on a single processor core, with an average docking time of 15 s per ligand, would take approximately 475 years. By contrast, VirtualFlow can dock 1 billion compounds in approximately 2 weeks when leveraging 10,000 CPU cores simultaneously. Such high-performance computing facilities are available to researchers through several potential sources, including computer clusters of local institutes, national super-computing centres or cloud computing platforms.

Targeting KEAP1 using VirtualFlow

To test the advantages of ultra-large-scale in silico screening and the performance of the VirtualFlow platform, we decided to target the challenging and therapeutically relevant protein–protein interaction between nuclear factor erythroid-derived 2-related factor 2 (NRF2) and Kelch-like ECH-associated protein 1 (KEAP1). NRF2 is a master regulator of cellular resistance to oxidative stress and cellular repair[7]. Under unstressed conditions, NRF2 is sequestered by KEAP1—an E3 ubiquitin ligase substrate adaptor—and targeted for degradation[8]. However, upon oxidative stress, reactive oxidants dissociate NRF2 from KEAP1 and NRF2 translocates to the nucleus to activate its transcriptional program of approximately 250 genes[9]. The NRF2–KEAP1 pathway is critical in protecting the cell under oxidative stress and inflammation and is implicated in a number of diseases[10]. There are ten drugs that target KEAP1 that are in clinical trials and nine more that are at the preclinical stage[10]. Using VirtualFlow, we screened approximately 1.3 billion compounds (around 1 billion compounds from the Enamine REAL library and about 330 million compounds from the ZINC library) against the NRF2 interaction interface on KEAP1. First, however, we describe the salient features of VirtualFlow and its scalability.

Characteristic features of VirtualFlow

One of the key features of VirtualFlow is its linear scaling behaviour ($O(N)$, where N is the number of cores) with respect to the number of CPUs and nodes used. VirtualFlow can run on computer clusters operated with any of the major resource managers (SLURM (https://slurm.schedmd.com), Moab/TORQUE (http://www.adaptivecomputing.com), PBS (http://www.pbspro.org), LSF (https://www.ibm.com/us-en/marketplace/hpc-workload-management) and SGE (http://gridscheduler.sourceforge.net)), and compatibility with additional job schedulers can be easily added. Thus VirtualFlow is also ideally configured for cloud computing platforms such as Amazon's Web Services (AWS), Microsoft's Azure and Google's Cloud Platform (GCP). VirtualFlow is able to run autonomously from the first to the last ligand in the screening pipeline, a feature that is facilitated by automatic submission of new batch system jobs. The workflow can be monitored and controlled during runtime. The VirtualFlow package consists of two applications that work seamlessly together: the VFLP (VirtuaFlow for Ligand Preparation) module—which prepares small molecules for screening—and the VFVS (VirtualFlow for Virtual Screening) module, which executes the virtual screening procedures (Fig. 1). The separation of ligand preparation and virtual screening is desirable because the same ready-to-dock ligand library can be used in any number of VFVS virtual screens.

The VFLP module

VFLP prepares ligand databases by converting them from the SMILES format to any desired target format (for example, the PDBQT format, which is required by many of the AutoDock-based docking programs). VFLP uses the JChem package of ChemAxon as well as Open Babel to desalt ligands, neutralize them, generate (one or multiple) tautomeric states, compute protonation states at specific pH values, calculate three-dimensional coordinates and convert the molecules into desired target formats (Extended Data Fig. 2). The output file formats that are currently supported by VFLP are shown in Supplementary Table 7.

Preparation of the Enamine REAL library

Commercially available compounds constitute the most interesting subset of the chemical space, as these compounds can be readily purchased. One of the largest vendor libraries that is currently available is the REAL library of Enamine, which contains approximately 1.4 billion make-on-demand compounds (as of October 2019, the ZINC15 database contained 1.46 billion compounds, but only provided 630 million molecules in a ready-to-dock format). We used VFLP to convert the approximately 1.4 billion compounds of the REAL library into PDBQT format (Methods) and have made this library freely available on the VirtualFlow homepage, accessible through a graphical interface (Supplementary Fig. 5). The entire database has a six-dimensional lattice architecture, the general concept of which was modelled after the ZINC15 database[11], in which each dimension corresponds to a physico-chemical property of the compounds (molecular mass, partition coefficient, number of hydrogen bond donors and acceptors, number of rotatable bonds and the topological polar surface area). The preparation of ligands using VFLP is a one-time effort.

The VFVS module

To set up a virtual screen with VFVS, a set of docking scenarios is specified by the user. Docking scenarios are defined by the choice of the external docking program, the receptor structure and the docking parameters (which include the pre-defined docking surface on the receptor, residues on the receptor that are allowed to be flexible during docking and the rigor of the docking routine). VirtualFlow currently supports the following docking programs: AutoDock Vina[12], QuickVina 2[13], Smina (which includes the Vinardo and AutoDock 4 scoring functions)[14], AutoDockFR[15], QuickVina-W[5], VinaXB[16] and Vina-Carb[17]. By supporting an array of different docking programs, VFVS can be used in a variety of cases by leveraging the unique advantages of each program. VFVS allows the specification of multiple docking scenarios to be carried out for each ligand, enabling consensus docking procedures, as well as ensemble docking procedures[18,19]. VirtualFlow is also amenable to the integration of other docking programs that are currently not a part of this platform.

Scaling behaviour of VFVS

To measure the scaling behaviour of VFVS, we measured the performance on two local clusters, LC1 and LC2. On LC1, we used 18,000 CPU cores of heterogeneous composition (different models of Intel Xeon and AMD Opteron processors), whereas on LC2 we used up to 30,000 Intel Xeon 8268 cores. The scaling behaviour was effectively linear in both cases (that is, $O(N)$) (Extended Data Fig. 3a). These results meet theoretical expectations as there is no direct communication between the processes running in parallel, which is key to perfect scaling behaviour without bounds. The independence of its parallel processes means that VirtualFlow is expected to scale linearly even if millions of cores are used. We also tested the performance of the platform on cloud-based computing systems, including GCP and AWS. On the GCP, we carried

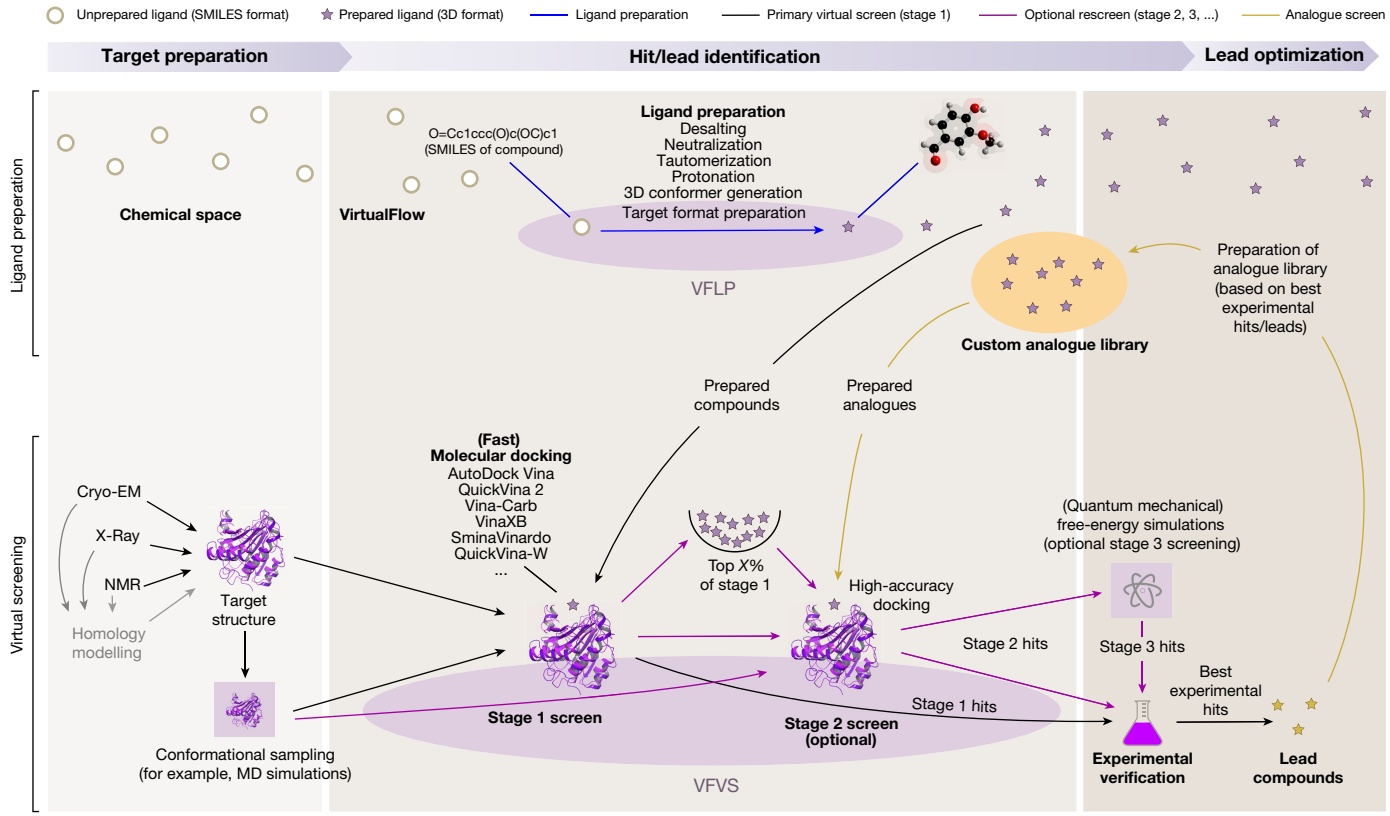


Fig. 1 | Application of VirtualFlow to the drug discovery process. Before the screening can begin, the target structure—which is generally obtained through X-ray, NMR, cryo-electron microscopy (cryo-EM) or homology-modelling studies—needs to be prepared. The preparation step can include molecular dynamics (MD) simulations to obtain one or more relevant conformations of the target protein. Once the structure is prepared, it can be used to identify new hit compounds by virtual screening-based approaches. The two independent modules of VirtualFlow, VFLP and VFVS, are designed to aid virtual screening-based needs. VFLP prepares (blue arrow) the desired chemical space into ready-to-dock ligand libraries, which can subsequently be used by VFVS during the virtual screen. The virtual screen usually consists of a primary virtual screen (stage 1), but can also be implemented using the multi-stage setting (pink arrows), which can include protein side chain flexibility or inclusion of multiple protein conformations (for example, from molecular dynamics simulations or NMR structures). In addition, complimentary software packages can be used during multi-staged screening (for example, to carry out quantum mechanics-based free-energy simulations as a final step) to improve the true hit rate and estimated binding affinities. After the virtual screening procedure, experimental verification can be carried out to identify true binders. Promising molecules (lead compounds) can be further optimized by creating custom analogue libraries and screening them again with VirtualFlow (yellow arrows), followed again by experimental verification of the hits.

out large-scale benchmarks with up to 160,000 CPUs and, despite this massive scaling in CPU volume, VirtualFlow still exhibited linear scaling behaviour (Extended Data Fig. 3a). A typical high-throughput screen, such as the one described in this study, of 1 billion compounds will take around 15 h on the GCP with 160,000 CPUs. The linear scaling behaviour over a large number of CPUs makes VirtualFlow suitable for the highly anticipated exascale computing age.

Multi-staged virtual screens with VFVS

VFVS can also be used to organize virtual screens with multiple stages to substantially increase the quality of the results (Fig. 2a). In the multi-staging approach, several virtual screens are executed in succession. The number of top-scoring compounds that advance from one stage to the next is successively reduced, with concomitant increases in docking accuracy and computational cost.

Using VFVS to screen 1.3 billion ligands

To validate the performance of VFVS, we screened a virtual library of 1.3 billion commercially available compounds (around 330 million compounds from the ZINC 15 database[11] and approximately 1 billion compounds from the Enamine REAL library) against KEAP1. It should be noted that there is some overlap of compounds between the two libraries.

This effort was completed in around 4 weeks, using on average approximately 8,000 cores on a heterogeneous Linux cluster.

To illustrate the benefit of an ultra-large-scale screen, we chose subsets of the ligands (0.1, 1, 10 and 100 million compounds) randomly from the around 1 billion compound screen of the REAL library and considered the scores of the top 50 compounds (Fig. 2b). As the scale of the screen increased, the average docking score increased, thus improving the chances of identifying molecules with higher binding affinity. This in turn leads to higher true hit rates and tighter experimental binding affinities, as predicted by a probabilistic model that we derived (Supplementary Information section D) and that has been experimentally demonstrated previously[2].

To demonstrate the use of VirtualFlow in a multi-staging context, we subjected the top approximately 3 million ranking compounds from the primary virtual screen to a rescoring procedure (Fig. 2a). In stage 2, the 13 residues of KEAP1 at the NRF2 interaction interface were allowed to be flexible. This flexibility accounted for the movement and/or dynamics of the amino acids at the binding interface, which are not captured by a static structure. In the rescoring procedure we used two different docking programs (Smina Vinardo and AutoDock Vina), and two replicas of each docking scenario were carried out to

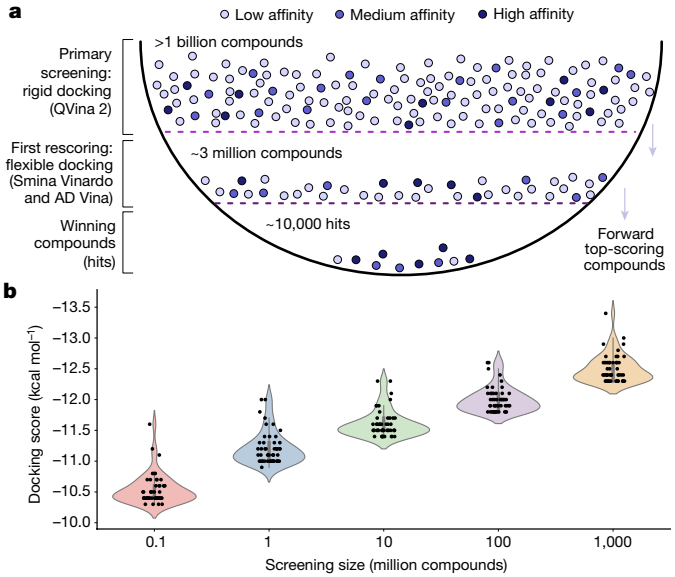


Fig. 2 | Schematic overview of the multi-stage screen and benefits of ultra-large-scale screens. a, In stage 1, approximately 1.3 billion compounds were screened with the fast docking program QuickVina 2 at the lowest accuracy level. In stage 2, 13 residues of the receptor were allowed to be flexible and the top approximately 3 million scoring compounds were rescored with higher accuracy using AutoDock Vina and Smina Vinardo. **b**, Violin plots of the docking scores of the top 50 molecules from virtual screens that targetted KEAP1 with different starting library sizes (0.1, 1, 10, 100 and 1,000 million ligands). To mimic virtual screens of smaller sizes, we randomly chose a subset of the ligands from the around 1 billion compound screen of the REAL library and considered the scores of the top 50 compounds. The docking score is an estimation of the free energy of binding (in kcal mol^{-1}) and, therefore, the more negative the value is, the tighter the hit binds to the target. The distributions show that the docking score of the top 50 compounds improves with the scale of the screen. The procedure was repeated independently five times with similar results.

further increase the conformational space sampled during the docking runs. The necessity of multi-stage screening depends on the target of choice and the computational resources available, but this type of virtual screen is particularly useful in cases in which dynamics at the docking interface is expected to have a marked role.

Experimental validation

From the in silico screen described above, we chose 590 hits for experimental validation. Of these, 492 compounds were from the top 0.03% of the stage 2 screen and 98 compounds were from the top 0.0001% of the stage 1 screen. Hits from stage 1 were ordered to compare the true hit rate between stage 1 and stage 2 hits, in a multi-stage setting. In addition to the ranking by docking score, the choice of these compounds was based on factors such as drug-likeness, availability for procurement, ligand efficiency and chemical diversity. We used four established biophysical methods: fluorescence polarization, surface plasmon resonance (SPR), nuclear magnetic resonance (NMR) and bio-layer interferometry (BLI) to experimentally validate the binding of the VirtualFlow-derived hits to KEAP1. Fluorescence polarization and SPR were initially used in a high-throughput manner (level 1) to detect binding and the compounds identified here were subsequently validated with more scrutiny in a detailed and low-throughput assay (level 2). We used a recombinantly expressed and purified Kelch domain of mouse KEAP1, henceforth referred to as KEAP1. An overview of the experimental verification workflow is graphically represented in Extended Data Fig. 6, and a detailed description of the experimental procedure is provided in the Methods. Of these four biophysical methods, fluorescence polarization and BLI detect the ability of the hits to displace the NRF2 peptide from KEAP1, identifying hits that we refer to as displacers. SPR and NMR directly detect binding of hits to KEAP1, identifying hits referred to as binders. VirtualFlow identifies molecules that potentially bind to the NRF2 interaction interface on KEAP1; however, the in silico screen is performed using KEAP1 alone, in the absence of NRF2. The NRF2-binding surface on KEAP1 is part of the deep pocket/tunnel of the KEAP1 β-barrel with NRF2 binding to the entrance of this tunnel. However, some compounds could bind more tightly by inserting deep into this central tunnel of KEAP1 rather than embracing the surface like the NRF2 peptide and/or bind to parts of KEAP1 that are not engaged by NRF2. Such binders might not effectively disrupt the interaction with NRF2, while still engaging KEAP1 with high affinity (Extended Data Fig. 9). In our experimental validation, we identified both displacers and binders.

Out of the cherry-picked 590 compounds, 69 were confirmed to bind to KEAP1 by level 2 SPR. To assess the ability of the compounds to displace the NRF2 peptide, we used the fluorescence polarization assay. Ten compounds were confirmed to be displacers with a half-maximum inhibitory concentration (IC_{50}) < 60 μM by fluorescence polarization and all of the compounds were identified as a binder by level 2 SPR. Interference by autofluorescence from the compounds themselves prevented the analysis of some of the compounds by fluorescence polarization. Thus, we used BLI as an orthogonal assay to assess the ability of the compounds to displace NRF2. The binding affinity of the NRF2 peptide to KEAP1 as measured by BLI was 1.86 nM, which is similar to that measured by fluorescence polarization, which was 3.67 nM (Extended Data Fig. 4). Of the 69 SPR level 2 active compounds, 40 compounds were able to disrupt the NRF2–KEAP1 interaction as observed by BLI. Of these 40 compounds, 16 were able to displace NRF2 from KEAP1 at a compound concentration of 20 μM, while all 40 compounds could do so at 100 μM. Using BLI, we were able to identify displacers that were missed by fluorescence polarization due to autofluorescence (an example is shown in Extended Data Fig. 8). We tested all of the SPR level 2 active compounds for potential aggregation by dynamic light scattering. We identified seven compounds that aggregated in the dynamic light scattering assay and hence were not considered for further evaluation (Supplementary Table 5). On the basis of the SPR level 2 and the fluorescence polarization level 2 binding data, we selected 23 compounds for SPR level 3 experiments to determine the binding affinity. All 23 compounds had affinities in the low micromolar to nanomolar range, and 12 compounds had submicromolar K_d values. From these 23 compounds, we tested the binding of 6 compounds (iKeap1, iKeap2, iKeap7, iKeap8, iKeap9 and iKeap22) to KEAP1 by a suite of NMR-based ligand-detection experiments. Out of these six compounds, five are displacers and one (iKeap9) is a binder. These six compounds were selected on the basis of the solubility constraints of the NMR experiments, the SPR K_d value and/or their ability to displace the peptide. We used differential line broadening, saturation transfer difference, Car–Purcell–Meiboom–Gill-based transverse relaxation time experiments, and protein-observed ^{1}H–^{13}C heteronuclear multiple-quantum correlation experiments to confirm the binding of the compounds to KEAP1. The ligand-detection NMR experiments confirmed that all six of the tested SPR level 3 active compounds bind to KEAP1 (Fig. 3 and Extended Data Figs. 7, 8). Protein-detection ^{1}H–^{13}C heteronuclear multiple-quantum correlation experiments show that the compounds engage KEAP1 in a specific manner, at the targeted NRF2-binding site. In the absence of resonance assignments, we use the fact that the compounds perturb a subset of KEAP1 resonances affected by the addition of the NRF2 peptide as evidence for competitive binding. These compounds are shown in Supplementary Figs. 1–3. Details about the other active compounds are provided in Supplementary Information section B.

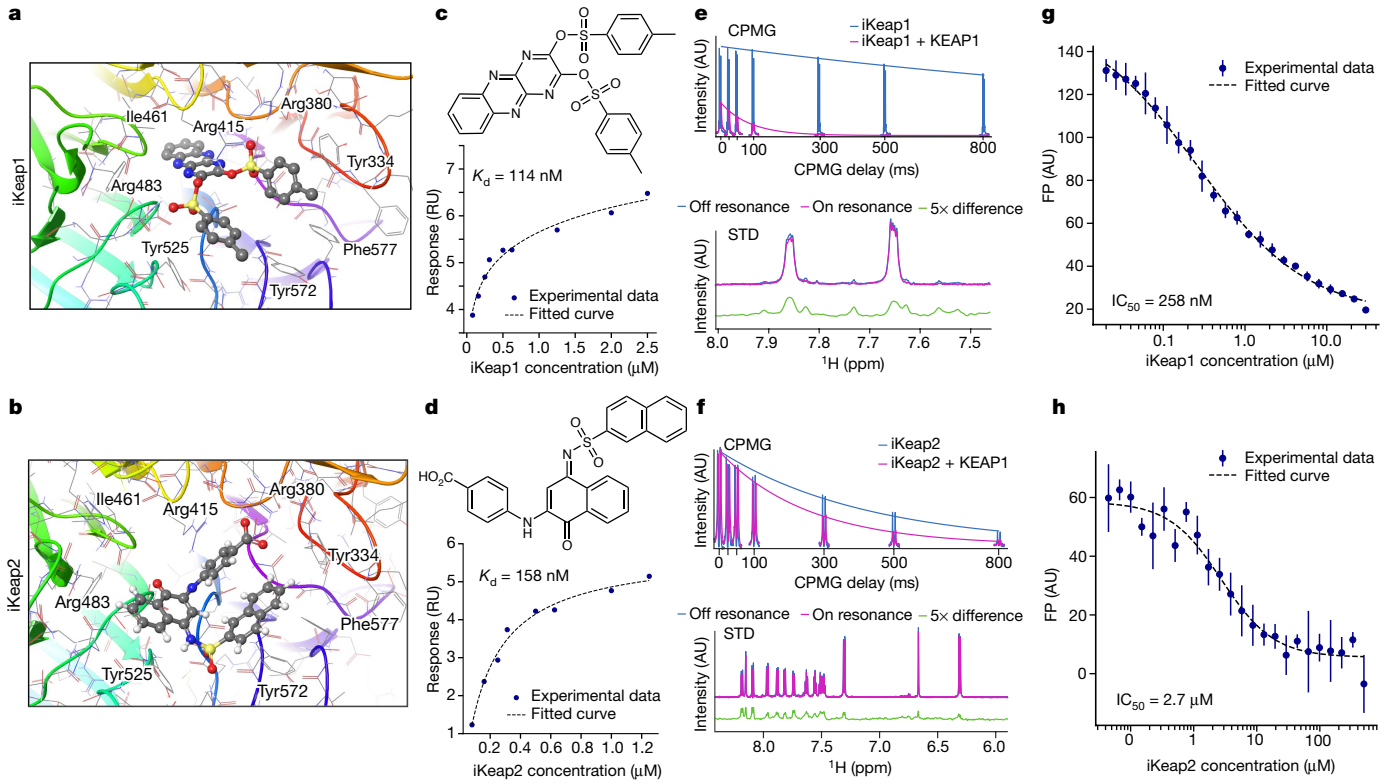


Fig. 3 | Docking poses and experimental verification of two hit compounds (iKeap1 and iKeap2). a, **b**, The docking poses of iKeap1 (**a**) and iKeap2 (**b**) were obtained from stage 2 of the virtual screening. **c**, **d**, SPR steady-state binding curves are shown for iKeap1 (**c**) and iKeap2 (**d**), showing clear binding with a K_d in the nanomolar range. We show one representative dataset from three independent experiments with similar results. RU, resonance units. **e**, **f**, Ligand-detection NMR experiments using Car–Purcell–Meiboom–Gill (CPMG)–R_2, which measures the transverse relaxation rate (R_2) of the ligand and saturation transfer difference (STD) NMR of iKeap1 (**e**) and iKeap2 (**f**) confirm the binding of the two compounds to KEAP1. AU, arbitrary units. **g**, **h**, Keap1 (**g**) and iKeap2 (**h**) were also functional in the fluorescence polarization (FP) assay, confirming that the compounds displace the peptide. The fluorescence polarization data shown here are from three technical replicates and the curve was fitted to the average value of the three technical replicates. Data are mean ± s.d. of the individual data points. The fluorescence polarization experiment was repeated independently twice with similar results and one representative result is shown here.

iKeap1 and iKeap2 are two of our top hits, and both these compounds are able to displace the NRF2 peptide from KEAP1. Both of the compounds are predicted to engage the NRF2-binding pocket of KEAP1, located at the entrance to the tunnel formed by the β-barrel (Fig. 3a, b). In comparison to iKeap2, iKeap1 descends deeper into this central tunnel of KEAP1. SPR results showed that iKeap1 and iKeap2 bind to KEAP1 with a binding affinity of 114 nM and 158 nM, respectively (Fig. 3c, d). NMR-based ligand-detection experiments confirmed that both iKeap1 and iKeap2 directly bind to KEAP1 (Fig. 3e, f). Fluorescence polarization assays showed that iKeap1 displaces NRF2 peptide with an IC_{50} of 258 nM and iKeap2 displaces the NRF2 peptide with an IC_{50} of 2.7 μM (Fig. 3g, h). BLI measurements additionally confirmed that both iKeap1 and iKeap2 are able to displace the NRF2 peptide from KEAP1. iKeap1 is similar to a previously reported naphthalene-based compound[20] with a lower IC_{50} (IC_{50} = 2.7 μM; compound C17 in Supplementary Table 1 and Extended Data Fig. 5d). C17 was identified as the best hit in a high-throughput screen of 270,000 compounds in a previous study[20].

Here, we also highlight iKeap7, which has the highest affinity as assayed by SPR (K_d = 15 nM) and displaces the NRF2 peptide with an IC_{50} of 38.2 μM (Extended Data Fig. 8). It should be noted that of the 14 hits described in the manuscript, only two hits, namely iKeap2 and iKeap7, contain pan-assay interference (PAINS) substructures. However, we performed a series of orthogonal binding assays, which confirmed that iKeap2 and iKeap7 are not experimental false positives. For details and discussion on how we verified that our experimental results were not affected by PAINS, see Supplementary Information section B.

Typically, protein–protein interactions have a larger interaction interface compared to the interface of the active site of an enzyme. Hence the in silico screen can identify binders that either partially overlap with the binding site of the interacting protein, such as iKeap9 (Extended Data Fig. 7) or those that bind in a manner that energetically favours the formation of the protein–protein complex. Examples of the latter, referred to as glues, have been previously described in the literature[21].

An open-source platform

To allow VirtualFlow to be used widely and develop dynamically, it is set up as a free and open-source project. GPU support is planned for the future and will be incorporated into VirtualFlow both natively and through external docking programs such as Gnina[22]. We encourage scientists to join the project and contribute to improving existing features, adding new features and functionality. The primary homepage of VirtualFlow, which provides additional resources, can be accessed at https://www.virtual-flow.org.

Outlook

VFVS can be used to search extremely large regions of the chemical space, which is the key to identifying promising small-molecule binders. VFVS is able to accomplish this by efficiently using high-performance computing resources, which will continue to increase in availability and power in the years to come, and novel virtual screening databases such

as the Chemical Universe Databases, which contain billions to trillions of compounds, that are still waiting to be explored[23].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2117-z.

1. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
2. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
3. Zhang, S., Kumar, K., Jiang, X., Wallqvist, A. & Reifman, J. DOVIS: an implementation for high-throughput virtual screening using AutoDock. *BMC Bioinformatics* **9**, 126 (2008).
4. Jiang, X., Kumar, K., Hu, X., Wallqvist, A. & Reifman, J. DOVIS 2.0: an efficient and easy to use parallel virtual screening tool based on AutoDock 4.0. *Chem. Cent. J.* **2**, 18 (2008).
5. Hassan, N. M., Alhossary, A. A., Mu, Y. & Kwoh, C.-K. Protein-ligand blind docking using QuickVina-W with inter-process spatio-temporal integration. *Sci. Rep.* **7**, 15451 (2017).
6. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
7. Yonchuk, J. G. et al. Characterization of the potent, selective Nrf2 activator, 3-(pyridin-3-ylsulfonyl)-5-(trifluoromethyl)-2*H*-chromen-2-one, in cellular and in vivo models of pulmonary oxidative stress. *J. Pharmacol. Exp. Ther.* **363**, 114–125 (2017).
8. Pallesen, J. S., Tran, K. T. & Bach, A. Non-covalent small-molecule Kelch-like ECH-associated protein 1-nuclear factor erythroid 2-related factor 2 (Keap1–Nrf2) inhibitors and their potential for targeting central nervous system diseases. *J. Med. Chem.* **61**, 8088–8103 (2018).
9. Davies, T. G. et al. Monoacidic inhibitors of the Kelch-like ECH-associated protein 1: nuclear factor erythroid 2-related factor 2 (KEAP1:NRF2) protein–protein interaction with high cell potency identified by fragment-based discovery. *J. Med. Chem.* **59**, 3991–4006 (2016).
10. Cuadrado, A. et al. Therapeutic targeting of the NRF2 and KEAP1 partnership in chronic diseases. *Nat. Rev. Drug Discov.* **18**, 295–317 (2019).
11. Sterling, T. & Irwin, J. J. ZINC 15—ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
12. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
13. Alhossary, A., Handoko, S. D., Mu, Y. & Kwoh, C.-K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics* **31**, 2214–2216 (2015).
14. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
15. Ravindranath, P. A., Forli, S., Goodsell, D. S., Olson, A. J. & Sanner, M. F. AutoDockFR: advances in protein-ligand docking with explicitly specified binding site flexibility. *PLOS Comput. Biol.* **11**, e1004586 (2015).
16. Koebel, M. R., Schmadeke, G., Posner, R. G. & Sirimulla, S. AutoDock VinaXB: implementation of XBSF, new empirical halogen bond scoring function, into AutoDock Vina. *J. Cheminform.* **8**, 27 (2016).
17. Nivedha, A. K., Thieker, D. F., Makeneni, S., Hu, H. & Woods, R. J. Vina-Carb: improving glycosidic angles during carbohydrate docking. *J. Chem. Theory Comput.* **12**, 892–901 (2016).
18. Amaro, R. E. et al. Ensemble docking in drug discovery. *Biophys. J.* **114**, 2271–2278 (2018).
19. Houston, D. R. & Walkinshaw, M. D. Consensus docking: improving the reliability of docking in a virtual screening context. *J. Chem. Inf. Model.* **53**, 384–390 (2013).
20. Marcotte, D. et al. Small molecules inhibit the interaction of Nrf2 and the Keap1 Kelch domain through a non-covalent mechanism. *Bioorg. Med. Chem.* **21**, 4011–4019 (2013).
21. Andrei, S. A. et al. Stabilization of protein–protein interactions in drug discovery. *Expert Opin. Drug Discov.* **12**, 925–940 (2017).
22. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein-ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
23. Reymond, J. L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Parallelization of the virtual screen using VirtualFlow

VirtualFlow uses four levels of parallelization in a hierarchical manner to enable it to run on batch system-managed Linux clusters of any configuration while allowing for perfect scaling behaviour. Each instance of VirtualFlow can submit multiple jobs, each job may use several job steps (currently only supported when using SLURM and Moab/TORQUE/PBS as the resource manager, whereas for SGE and LSF only single job steps per job are possible), one job step is able to execute an arbitrary number of queues, and each queue executes the external programs that are processing the ligands (Extended Data Fig. 1). These programs may be additionally parallelized internally, for instance through multithreading. Details about the workflow within a single queue are provided in Supplementary Information section G.

Workload balancing

When processing ligands in parallel, there needs to be a mechanism that makes sure that each ligand is treated only once. However, one main problem with parallelization is that most cluster file systems are too slow to work off a single simple task list. This is because, when different processes access the file at the same time, clashes can occur, as it may take up to several seconds until one job sees the changes made to a file by another job. These latency problems also mean that file locking mechanisms do not prevent these clashes. The standard solution for solving this kind of problem is to let different processes communicate directly with each other or through a central master process. However, in most cases, this results in sub-linear scaling behaviour, which normally worsens as more and more parallel running processes become involved. Moreover, many advanced parallelization methods such as MPI or OpenMP, do not allow for inter-job communication, while in many cases multiple simultaneously running jobs are needed. Therefore, to maintain perfect scaling behaviour that allows multiple jobs and a virtually unrestricted number of CPUs, we have developed an advanced task-list mechanism. The key is to minimize the number of instances that the parallel processes need to access the task list. The mechanism that we have implemented requires only a single access per batch system job, each of which can contain a large number of parallel running processes. For this purpose, we have implemented a workload balancer, which distributes the tasks from the central task list at the beginning of each job to all the queues that belong to it. The central task list contains collections of ligands as elementary components (rather than individual ligands), and the workload balancer takes into account the length of each collection when distributing them among the queues. This approach markedly reduces the number of times the central task list has to be accessed. For example, if the workflow uses 10 jobs in parallel, and each job runs on 100 nodes with a wall time (real run time) of 1 week and 24 CPU cores per node, and one ligand requires approximately 30 s to be docked, then the central task list needs to be accessed only 10 times per week to feed a total of 24,000 parallel running queues (assuming each queue runs on one CPU core). In this case, approximately 483,840,000 ligands are processed in 1 week, which means that the advanced task list approach reduced the number of accesses to the central task list by a factor of 48,384,000 in comparison to the number needed by a trivial task list approach (one access per ligand processed). This factor can be improved even further depending on the job size and the cluster wall time. In case two parallel processes want to access the central task list simultaneously, two backup mechanisms were implemented. The first mechanism is a time-dispersion mechanism, which spreads out simultaneously arriving jobs in time, and further stalls subsequent jobs until the workload balancer of the current job is finished. If this mechanism should fail to prevent a simultaneous access event, which could result in a damaged or empty task file, a second mechanism restores the task list using an automatically backed-up copy of a previous version of the central task list. More information about the input and output file structures of VirtualFlow can be found in Supplementary Information section F.

Reduction of input and output load

One of the potential bottlenecks of computer clusters is the input and output load that they can handle, even when they utilize shared cluster file systems with high bandwidth. The limit of the input and output capacities of a cluster can be easily reached if many small processes that individually handle their input and output and use the shared file system are running in parallel. This circumstance can pose a serious problem when running large-scale workflows with thousands of queues working in parallel, and can easily lead to crashing the cluster file system. To address this problem and considerably minimize the load on the shared file system, VirtualFlow is able to perform most input and output operations on the local temporary file systems of the computing nodes, which are normally fast RAM-based (virtual) drives readily available on any Linux system (usually /dev/shm). The final output files are then stored in batches at large time intervals on the permanent cluster file system.

Preparation of the ligand databases

One of the ligand databases that was screened originates from the state of the ZINC 15 database in the November of 2016. Approximately 330 million compounds were downloaded in SMILES format and converted into three-dimensional PDBQT files with VFLP because, at the time, the ZINC 15 database only provided a fraction of the compounds in a ready-to-dock format. During the conversion, the molecules were protonated with cxcalc (ChemAxon) and the three-dimensional structure of the ligand was computed by the molconvert tool of ChemAxon (https://chemaxon.com/products/jchem-engines). If protonation or the generation of the three-dimensional structure failed, Open Babel[24] was used as a fall-back option. Other preparation steps, such as desalting, were not carried out on these compounds as they had already undergone these basic preparation steps for the ZINC 15 database.

We also prepared the compounds in the REAL database provided by Enamine (https://enamine.net/library-synthesis/real-compounds). Approximately 700 million partially stereospecific SMILES were expanded into fully stereospecific SMILES, resulting in around 1.4 billion molecules. These were then prepared with VFLP into a ready-to-dock format. Specifically, the compounds were desalted and neutralized using cxcalc (ChemAxon), major tautomers were computed using cxcalc and then protonated using cxcalc (using Open Babel as fall back), the three-dimensional coordinates were computed using molconvert (ChemAxon) (using Open Babel as a fall back) and, finally, compounds were converted into the PDBQT format using Open Babel. This library has been made available through an interactive web interface (Supplementenary Information section C). The scaling behaviour of VFLP was measured on the GCP using up to 20,000 CPU cores and these data are shown in Supplementary Fig. 7.

Computation time of VFVS

The total computation time (T) is directly proportional to the number of ligands screened (N) and the processing time per ligand (P), and inversely proportional to the number of CPUs used (C):

$$T \propto (P \times N)/C$$

The processing time per ligand (P) depends mainly on the specific docking scenario (which includes the receptor and all of the possible docking options and parameters) and the speed of the CPUs used, and can be approximated by:

$$P \propto (E \times \Theta + \zeta)/\eta$$

where η is a factor that represents the CPU speed relative to a reference CPU, E is the docking exhaustiveness parameter (elaborated in 'Relationship between the exhaustiveness parameter and the docking time'), Θ is the docking time per unit exhaustiveness on the reference CPU (that is, the slope of the lines shown in Extended Data Fig. 3c) and ζ is the initial set-up time required by the docking program on the reference CPU (that is, the intersection at the y axis of the lines in Extended Data Fig. 3c). For a typical case of a large-scale first-stage virtual screen on one of the newer Intel CPUs, the average processing time per ligand (P) is roughly 5 s using the fastest docking settings. It follows that when 5,000 CPUs are used, the total screening time for 100 million compounds will be roughly 30 h. Extended Data Figure 3b illustrates the relationship between the computation time and the number of CPUs for a given number of ligands (assuming an average processing time of 5 s per compound).

Relationship between the exhaustiveness parameter and the docking time

The time to dock a single molecule depends on the number of conformations that are sampled, and this number is largely independent of the size of the docking box or surface area. The number of conformations sampled can be controlled by the exhaustiveness parameter of the docking programs. Docking time has a linear dependency on the exhaustiveness parameter, as shown in Extended Data Fig. 3c. The inset in the graph shows the slope for each of the docking programs, providing an estimate of the degree of dependency between the computational time and the exhaustiveness parameter for individual docking programs.

As most docking programs use a probabilistic search algorithm, the results of separate iterations with the same starting set-up can differ. This can be beneficial as it can be more efficient to carry out multiple less-exhaustive docking iterations than to run one highly exhaustive iteration. The exhaustiveness here is a measure of the extent to which the conformational space of the ligand, and potentially the protein side chains, is explored by the search algorithm during the docking procedure. In light of this, VFVS can be configured to carry out multiple replicas per docking scenario, thus improving the overall efficiency.

Lead optimization using VFVS

The operational flexibility enables VFVS to also be used during lead optimization (Fig. 1). In this context, a library of analogues of a chosen lead compound can be prepared with VFLP and screened by VFVS with high docking accuracy (for example, setting the exhaustive parameter to a high value, allowing specific amino acids in the binding interface to be flexible, using multiple docking programs and/or multiple receptor (backbone) conformations), which can considerably accelerate the lead optimization process.

Parameters of the virtual screen against the KEAP1 target

For the virtual screening validation test, the crystal structure of the KEAP1 Kelch domain (Protein Data Bank (PDB) ID: 5FNQ)[9] was used.

The protein was stripped of all small molecules present (including water), protonated at physiological pH and then converted into PDBQT format using AutoDockTools[25].

The NRF2-binding interface on KEAP1 was chosen as the target of the screening and the exact location was determined using previously published co-crystal structures of KEAP1 and the NRF2 peptide (PDB ID: 4IFL). The in silico screen was carried out as follows. First, VFVS used the docking program QuickVina 2 in an initial (primary) virtual screen with mouse KEAP1 as a rigid receptor structure.

In this primary virtual screening, the docking search space was a rectangular parallelepiped (that is, a cuboid) of size 15.0 × 16.5 × 14.275 Å^3. The exhaustiveness parameter was set to 1, which favours fast computational times. The quality of individual docking results, and therefore the ranking, depends largely on the external docking program chosen (which is independent of VirtualFlow).

Next, in the rescoring procedure, the following amino acid side chains at the binding interface were allowed to be flexible: Tyr334, Arg380, Asn382, Arg415, Cys434, His436, Ile461, Phe478, Arg483, Ser508, Tyr525, Tyr572 and Phe577. AutoDockTools was used to generate the rigid and flexible receptor structures in PDBQT format. The exhaustiveness was set to 1, and two replicas (iterations) were carried out of each docking scenario (with Smina Vinardo and AutoDock Vina as the docking programs). The size of the docking box was set to 27.0 × 27.0 × 24.0 Å^3.

Expression and purification of GST–KEAP1

A codon-optimized sequence of the Kelch domain (residues 322–624) of mouse KEAP1 cloned into a pGEX-6P-3 vector with BamHI and XhoI cloning sites, and an NRF2 peptide (AFFAQLQLDEETGEFL) with an N-terminal tetramethylrhodamine (TAMRA) fluorophore were purchased from GenScript USA. The pGEX-6P-3 vector contains an N-terminal glutathione *S*-transferase (GST) tag, which is expressed as a fusion with the target sequence, resulting in a gene product that will henceforth be referred to as GST–KEAP1. The vector carrying GST–KEAP1 was transformed into BL21(DE3) *Escherichia coli*. The transformed cells were grown at 37 °C to an optical density of 0.6 at a measurement wavelength of 600 nm and protein expression was induced with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). The cells were allowed to grow for 12–16 h at 18 °C and were subsequently collected by centrifugation at 4,200 rpm for 20 min at 4 °C.

To purify GST–KEAP1, cell pellets from 2 l of culture were resuspended in 40 ml of GST-binding buffer (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM EDTA) supplemented with 3.5 mM β-mercaptoethanol and protease inhibitors (Roche). Cells were lysed by sonication, the insoluble fraction was removed by centrifugation at 16,000 rpm and the soluble fraction was applied to 10 ml of GST agarose resin (GoldBio). The suspension was nutated for 4 h at 4 °C, and the unbound fraction was removed by gravity-flow chromatography. The slurry was washed twice with GST-binding buffer supplemented with 3.5 mM β-mercaptoethanol. The bound fraction was eluted from the slurry with 20 mM reduced glutathione in GST-binding buffer. The resulting eluate was loaded on a Superdex 200 size-exclusion chromatography (SEC) column pre-equilibrated in SEC buffer (20 mM Tris-HCl, pH 8.0, 50 mM NaCl, 10 mM dithiothreitol).

Fluorescence polarization assays

Measuring the dissociation constant of the NRF2-KEAP1 interaction. We prepared 2 nM TAMRA–NRF2 peptide in fluorescence polarization buffer (20 mM Tris-HCl pH 8.0, 50 mM NaCl, 10 mM DTT, 2 mM 3-[(3-cholamidopropyl)-dimethylammonio]-1-propanesulfonate (CHAPS), 0.005% BSA, 1% DMSO) in 384-well plates (Corning, 3575), to establish the K_d of the interaction between TAMRA–NRF2 and GST–KEAP1. GST–KEAP1 was titrated into the TAMRA–NRF2 peptide starting at a concentration of 76 μM GST–KEAP1 followed by twofold dilutions for a total of 24 points. A K_d of 3.67 ± 0.35 nM was determined for the interaction (Extended Data Fig. 4a).

Fluorescence polarization, level 1 (high-throughput screening of VirtualFlow derived hits). All 590 compounds ordered for testing were dissolved in DMSO-d_6 to a final concentration of 10 mM. Two AB1056 (Abgene) plates were prepared as source plates for screening. The first source plate contained 11 μl of each of the 10-mM compounds. The second source plate was filled with 9 μl DMSO and 1 μl from the first source plate was transferred into the second via pin transfer with a Vprep liquid-handling pipetting station (Agilent), resulting in a final concentration of 1 mM for each compound in the second source plate. Then, 384-well (Corning, 3575) assay plates were pre-loaded with 7 nM GST–KEAP1 in fluorescence polarization buffer (30 μl per

well). Subsequently, 300 nl, 100 nl and 33 nl volumes were transferred from each source plate (the 10-mM and 1-mM plates) to pre-loaded 384-well assay plates. The assay plates were incubated for 1 h at room temperature before 2 nM TAMRA–NRF2 peptide was added to each well with an HP D300 (Hewlett-Packard). After 3 h of incubation at room temperature, fluorescence polarization (excitation, 485 nm; emission, 520 nm) was measured using an EnVision plate reader (PerkinElmer). This assay resulted in six-point titrations, which are not sufficient to calculate accurate IC_{50} values, but enable for the selection of top binders.

Fluorescence polarization, level 2 (screening of top hits). The 27 compounds that were active in the fluorescence polarization level 1 assay were subjected to a second 24-point fluorescence polarization screen (level 2), starting from 500 μM compound followed by 1.5-fold serial dilution. For the best compound, iKeap1, the starting concentration was reduced to 30 μM and the following concentrations were used in the titration: 30.00 μM, 21.60 μM, 15.50 μM, 11.10 μM, 8.00 μM, 5.76 μM, 4.14 μM, 2.98 μM, 2.15 μM, 1.54 μM, 1.11 μM, 0.80 μM, 0.576 μM, 0.414 μM, 0.298 μM, 0.215 μM, 0.154 μM, 0.111 μM, 0.080 μM, 0.0606 μM, 0.0459 μM, 0.0348 μM, 0.0264 μM and 0.02 μM. The measurements were carried out in triplicates, and the three data points for each concentration were averaged. IC_{50} values were determined by fitting the averaged data points to a four-parameter logistic curve using the nonlinear least-squares method provided by the SciPy library for Python (https://www.scipy.org/). The standard error (Supplementary Table 4) of the IC_{50} was computed by taking the square root of the diagonal of the parameter covariant matrix.

BLI assays

BLI binding and displacement assays. NRF2–KEAP1-binding BLI experiments were performed on an Octet RED384 (ForteBio) using streptavidin-coated Dip and Read Biosensors (ForteBio) and 384-well plates with 120 μl volume. The sensors were incubated for 5 min in 500 nM biotinylated NRF2 peptide in binding buffer (10 mM HEPES, pH 7.5, 50 mM NaCl, 0.1% (v/v) Tween-20 with 0.5 mM TCEP and 1% DMSO). To test for nonspecific binding of the GST–KEAP1 protein, reference tips were incubated in buffer only. The tips were washed with buffer for 2 min to obtain a baseline reading and then transferred to wells containing various concentrations of GST–KEAP1 protein (100 nM, 50 nM, 25 nM, 12.5 nM, 6.75 nM, 3.375 nM, 1.679 nM and 0.844 nM) for 10 min. After measuring association, tips were moved to wells containing buffer, and dissociation was measured for 5 min. The data were processed and analysed using the Octet data analysis software version 11.0 (ForteBio). The association–dissociation curve for each concentration was fitted using a 1:1 model given by the equations:

$$R_t^{\text{on}} = \frac{k_{\text{on}} \times C}{k_{\text{on}} \times C + k_{\text{on}}} R_{\text{max}} (1 - e^{-(k_{\text{on}} \times C + k_{\text{off}}) \times t})$$

$$R_t^{\text{off}} = R_{\text{eq}} \times e^{-k_{\text{off}} \times t}$$

where R_t^{on} and R_t^{off} are the BLI signals at time t, R_{eq} is the equilibrium response, k_{on} is the association rate constant, k_{off} is the dissociation rate constant, C is the analyte (protein) concentration and R_{eq} is the signal level at the equilibrium of association that depends on the analyte (protein) concentration and the maximal capacity (R_{max}) of the sensor surface. By computing the ratio $k_{\text{off}}/k_{\text{on}}$, the apparent equilibrium constant K_{d} is obtained. The resulting apparent K_{d} values were averaged.

Compound screening by BLI displacement assay. The BLI displacement assays were set up as described above. The biotinylated NRF2 peptide was used at a concentration of 500 nM and GST–KEAP1 protein was used at a concentration of 25 nM. The compounds were used at concentrations of 20 and 100 μM, and pre-incubated with GST–KEAP1 protein. The association phase was measured in the well containing compound with GST–KEAP1 protein for 10 min, and followed by a dissociation phase in buffer for 5 min. The inhibition percentage was the average BLI signal in the last 50 s of the dissociation phase, normalized against the condition of GST–KEAP1 protein in the absence of compound. The dose-dependent experiment with iKeap22 was carried out at 10 μM, 20 μM, 40 μM, 80 μM and 100 μM compound concentration and pre-incubated with 25 nM GST–KEAP1 protein.

To test for nonspecific binding of the compounds, the sensor was coupled with biotinylated NRF2 peptide and the compounds were used at 20 μM concentration without protein.

SPR binding assays

All SPR binding experiments were performed on a BiacoreT200 (GE Healthcare) instrument at 25 °C in running buffer (10 mM HEPES pH 7.5, 50 mM NaCl, 0.1% (v/v) Tween-20 with or without 0.5 mM TCEP and 1% DMSO). The running buffer was prepared fresh on each day of use, filtered and degassed before the SPR experiments. The target protein (GST–KEAP1) was anchored to a CM5 chip using a GST labelling kit (GE Healthcare)[26], in which a polyclonal goat anti-GST antibody was immobilized on a CM5 sensor chip by amine-coupling method 1.

SPR assay, level 1 (one-point high-throughput screen). The SPR level 1 screening was carried out as previously reported[27]. First, we prepared 10 mM DMSO-d_6 stock solutions of the 590 compounds that were procured in powder form. Then, 20 μM samples of the compounds were made by diluting the stock compounds in running buffer with 0.5 mM TCEP and 1% DMSO. The anti-GST immobilizing chip was saturated with GST in the reference channel and GST–KEAP1 in the target channel with resonance unit (RU) values of 750–800 for GST and 2,000–3,000 for GST–KEAP1. Binding of compounds to the immobilized protein was monitored for 60 s in both the association and dissociation phase. Additional injection of the running buffer was performed after every compound binding. All binding signals (RU_{max} = 16–29 RU, 1:1 stoichiometry) were corrected for the signals from the reference channel and buffer blank. Compounds were classified as an SPR level 1 hit if the condition RU > 4 was satisfied. This criterion was based on the positive control (iKeap1, RU = 4.65 ± 0.74).

SPR assay, level 2 (five-point high-throughput screen of the SPR level 1 hits). The hits from the SPR level 1 assay were rescreened at five different compound concentrations (0.5, 1, 5, 10 and 20 μM), in running buffer with 0.5 mM TCEP and 1% DMSO, at a rate of 30 μl min^{-1}. The hits were classified as hits if they produced a concentration-dependent SPR response and RU > 4 at a compound concentration of 20 μM.

SPR assay, level 3 (SPR experiments of selected SPR level 2 hits). We chose 23 out of the 69 SPR level 2 hits for level 3 analysis. Given the low throughput of the level 3 SPR assay, we chose a subset of the SPR level 2 hits, which included the displacers from the level 3 fluorescence polarization assay, the compounds that were tested by NMR and select SPR level 2 hits. SPR experiments were carried out in which the target protein (GST–KEAP1) was captured and regenerated in each compound cycle. All SPR data processing and analyses were performed using the BIAevaluation software (version 3.0). For steady-state binding, the R_{eq} signal was plotted against the analyte concentration and fitted to the one-site or the biphasic binding model (Supplementary Table 3) using the Levenberg–Marquardt algorithm used by the BIAevaluation software. The one-site binding model is given by the equation

$$R_{\text{eq}} = (R_{\text{max}} \times C)/(K_{\text{d}} + C) + b \quad (1)$$

where R_{eq} is the SPR signal at equilibrium, R_{max} is the SPR signal at saturation of the binding mode, K_{d} is the dissociation constant of the compound, b is the offset and C is the concentration of the compound. The biphasic binding model is given by the equation

$$R_{eq} = (R_{max,1} \times C)/(K_{d,1} + C) + R_{max,2} \times C/(K_{d,2} + C) + b \quad (2)$$

where R_{eq} is the SPR signal at equilibrium, $R_{max,1}$ and $R_{max,2}$ are the SPR signals at saturation of the two binding modes, $K_{d,1}$ and $K_{d,2}$ are the dissociation constants of the compound corresponding to the two binding modes, b is the offset and C is the concentration of the compound.

Standard errors of the estimated K_d values were computed with the BIAevaluation software, which computes them using the diagonal elements of the covariance matrix and the residual. The software operates based on the equations found on page 378 of the book *Receptor-Ligand Interactions: A Practical Approach*[28].

Ligand-detection NMR experiments

The differential line broadening experiments serve as simple one-dimensional experiments, in which the proton signal of the ligand is monitored. The ligand concentration exceeds the receptor concentration (for example, 10–20-fold) in this experiment and broadening of the resonance frequencies in the presence of the receptor is a consequence of ligand molecules shuttling between free and bound states. Differential line broadening manifests as a broadening of the ligand resonance due to binding of a protein. The ligand is in equilibrium between the free and protein-bound states dictated by the equilibrium constant. Differential line broadening is the result of the change in relaxation rate and the difference in chemical shift of the bound ligand. In the saturation transfer difference (STD) experiments, a region of the spectral space (−1 to 0.5 ppm) that has resonances from the receptor but not the ligand is selectively saturated. Resonances from methyl-bearing amino acids (Ile, Leu and Val) often populate this region of the spectral space. This saturation is transferred to the rest of the protein and eventually to the bound ligand by spin diffusion. In the implementation of STD, two one-dimensional spectra are recorded in an interleaved fashion. In the first experiment neither the receptor nor the ligand is saturated (off-resonance) and in the second the receptor is selectively saturated (on-resonance). Spectra of free ligands are observed in both experiments. However, if the ligand transiently binds to the receptor then the saturated receptor will transfer magnetization to the ligand. This transfer will be reflected as reduced intensity in the on-resonance saturated spectrum compared to the off-resonance saturation. The results are often presented as a difference spectrum between the on- and off-resonance saturation experiments. The appearance of ligand resonances in the difference spectrum is indicative of ligand binding. Measurement of the transverse relaxation rate of the ligand is another complementary strategy to detect ligand binding to a receptor. The free ligand behaves like a small molecule and experiences slow transverse relaxation; however, transient binding to the receptor enhances the transverse relaxation rate of the ligand. Thus, an increased transverse relaxation rate in the presence of a receptor directly indicates binding to the receptor. In the experimental set-up, a series of one-dimensional experiments in which the coherences of the ligand spend increasing amounts of time in the transverse plane is recorded. Ligands that engage the protein will relax faster than unbound ligands. We refer to these experiments as Car–Purcell–Meiboom–Gill (CPMG) or CPMG–R_2 experiments. Although any of these experiments are in principle sufficient to demonstrate ligand binding, false positives for either of these experiments have been reported. However, a detection of a false-positive hit is highly unlikely if all three experiments indicate binding, which is the case for all of the hits reported here.

All of the ligand-detecting experiments were performed with 50 μM compound alone or in the presence of 5 μM KEAP1 (without the GST tag) in NMR buffer (phosphate-buffered saline supplemented with 5% DMSO-d_6 and 4 mM deuterated DTT at pH 7.4) unless otherwise noted. For iKeap1 and iKeap2, the protein concentration was kept at 2.5 μM due to tight binding. ^{1}H one-dimensional spectra of the compounds were recorded in the absence and presence of KEAP1 to assess line broadening. STD spectra of the compounds in the presence of KEAP1 were recorded with 3 s saturation time on (0 ppm) and off (−20 ppm) resonance, respectively. The relaxation rate of the compounds was measured in the absence and presence of KEAP1 with a series of ^{1}H one-dimensional experiments with CPMG-based transverse relaxation time filters of various lengths: 1 ms, 25 ms, 50 ms, 100 ms, 300 ms, 500 ms and 800 ms. Data were analysed and visualized in MATLAB (MathWorks).

Protein-detection NMR experiments

The cleaved Kelch domain (residues 322–624) of mouse KEAP1 consists of 308 amino acids with close to 300 detectable amide resonances. Therefore, correlating chemical shift perturbations of small-molecule inhibitors to perturbations introduced by NRF2 would have been prohibitively difficult in ^{1}H–^{15}N HSQC spectra without full backbone assignment. Our aim was to rely on methodology that can be quickly and easily implemented even for very large proteins for which backbone assignment might not be feasible. Indeed, with a molecular mass of 33.7 kDa, the Kelch domain (322–624) of mouse KEAP1 is already on the larger side for NMR backbone assignment. To overcome the spectral crowding in a ^{1}H–^{15}N HSQC spectrum and minimize problems due to low ligand solubility, we implemented a ^{1}H–^{13}C TROSY heteronuclear multiple-quantum correlation experiment coupled with fast data acquisition. For the protein-detected ^{1}H–^{13}C heteronuclear multiple-quantum correlation experiments, a sample of KEAP1 that is selectively labelled with ^{1}H and ^{13}C at the methyl groups of isoleucine, leucine and valine (ILV) residues, in an otherwise deuterated background, was used. This labelling strategy is referred to as ILV labelling. ILV-labelled samples of KEAP1 were prepared by culturing BL21(DE3) *E. coli* cells containing a plasmid for GST–KEAP1, in perdeuterated M9 medium with 1 g ^{15}N-NH_4Cl and 2 g ^{2}H-^{12}C-glucose in ^2H_2O. Then, 1 h before induction with IPTG, 330 mg l^{-1} 2-(^{13}C)methyl-4-($^2H_3$)-acetolactate (a precursor for leucine and valine) was added. The acetolactate was activated beforehand as previously described[29]. Subsequently, 20 min before induction, 75 mg l^{-1} ^{13}C/^{1}H-methyl ketobutyrate sodium (a precursor for isoleucine), which was otherwise deuterated, was added. The use of acetolactate resulted in stereospecific ^{1}H–^{13}C labelling of only one of the leucine(λ2) and valine(γ2) methyl groups as previously described[29]. The protein was purified as described above. The GST-tag was cleaved by preScission protease cleavage and the free Kelch domain of mouse KEAP1 was eluted in NMR buffer from a SEC column. All NMR measurements for the ILV-labelled KEAP1 were performed at a protein concentration of 5 μM. The protein concentration was kept low to account for poor solubility (for NMR) of some of the compounds. The concentrations of the compounds were 50 μM, except for iKeap1 and iKeap2, for which the concentrations were 25 μM, owing to poor solubility.

Given the low concentration of the protein, we used the methyl SOFAST methyl TROSY with 46 ms and 18 ms acquisition times in the direct and indirect dimensions, respectively[30]. The spectral width was set to 14 ppm (^{1}H) and 20 ppm (^{13}C) in the direct and indirect dimensions, respectively, and the spectrum was recorded at 298 K on an 800-MHz Bruker spectrometer equipped with an AVANCE III console and a cryogenically cooled probe. A 4.5-ms Pc9_4_90.1000 pulse was used for selective excitation of the methyl ^{1}H resonances and a 1.2-ms Rsnob.1000 pulse was used to selectively refocus proton chemical shift evolution and ^{1}H–^{13}C J-coupling during ^{13}C chemical shift evolution. Proper choice and calibration of the excitation and refocusing pulses is crucial to avoid perturbing the water signal, which can significantly lower the achievable signal to noise ratio. Fast data acquisition was achieved with a 150-ms recycling delay, which enabled the recording of experiments with 512 scans in 5 h.

Detecting aggregation using dynamic light scattering

To test the potential aggregation of hits, we used dynamic light scattering experiments. The experiments were performed on a ZS90 Zetasizer instrument (Malvern Panalytical). Measurements were done in

triplicate with 10 scans per run (100 s). The compounds were used at 20 μM concentration in running buffer (10 mM HEPES pH 7.5, 50 mM NaCl, 0.1% (v/v) Tween-20 with or without 0.5 mM TCEP, 2% DMSO), which was filtered before use. The 20-μM working solution was made from a 1 mM stock of the compound in DMSO. The data were analysed by the built-in software. Compounds were classified as aggregated when the radius of the measured particles was above the minimum colloidal aggregate size (for small molecules)[31] of 50 nm.

In addition, the solubility of iKeap1, our most potent displacer, was analysed with an NMR solubility assay based on a previously described technique[32]. We made individual samples of iKeap1 at various concentrations (in PBS buffer, pH 7.4) ranging from 5 μM to 30 μM and measured the one-dimensional NMR spectrum of each sample with identical experimental conditions. The resonances of iKeap1 were then integrated and plotted as a function of the concentration. The plot shows a linear trend ($R^2 = 0.996$), indicating that iKeap1 does not aggregate in this concentration range (Supplementary Fig. 4).

Excluding interference from PAINS

PAINS comprise 480 markers initially identified as moieties postulated to cause interference in experimental high-throughput screens[33]. PAINS are often found in the databases that are commonly used for in silico screens, and the user should be cognizant of the fact that a potential hit could contain a PAINS substructure. However, it should also be noted that certain PAINS-like aspects can be mitigated by judicious use of medicinal chemistry, and some aspects of PAINS could have no effect, depending on the target of choice and/or the experimental assays used[34,35]. Attention should be paid to identifying and rigorously characterizing any PAINS among the hits identified in an in silico screen.

Two of the hit compounds (iKeap2 and iKeap7) reported in this manuscript contain PAINS substructures. We performed additional experiments to confirm that iKeap2 and iKeap7 are not false positives due to assay interference. First, we used dynamic light scattering to confirm that all of the compounds shown here did not aggregate at the concentrations used in the various experiments (Supplementary Table 5). Second, we carried out ligand-detection NMR experiments using STD NMR and CPMG performed with a tenfold excess of the compound to show that iKeap2 and iKeap7 bind KEAP1 in a reversible manner (Fig. 3 and Extended Data Fig. 8). And finally, we carried out protein-observed ^{1}H–^{13}C heteronuclear multiple-quantum correlation experiments to show that both iKeap2 and iKeap7 engaged KEAP1 in a specific manner at the NRF2-binding site and did not aggregate the protein (Supplementary Figs. 1, 3). In the event that these compounds caused KEAP1 to aggregate, all of the resonances would be broadened, which is not the case here.

Statistics and reproducibility

Characterization of the violin plots in Fig. 2b were as follows. Screening size of 100,000: minimum, −10.3 kcal mol^{-1}; maximum, −11.6 kcal mol^{-1}; median, −10.4 kcal mol^{-1}; first quartile (Q_1), −10.4 kcal mol^{-1}; third quartile (Q_3), −10.6 kcal mol^{-1}. Screening size of 1 million: minimum, −10.9 kcal mol^{-1}; maximum, −12 kcal mol^{-1}; median, −11 kcal mol^{-1}; Q_1, −11.1 kcal mol^{-1}; Q_3, −11.3 kcal mol^{-1}. Screening size of 10 million: minimum, −11.675 kcal mol^{-1}; maximum, −12.3 kcal mol^{-1}; median, −11.5 kcal mol^{-1}; Q_1, −11.4 kcal mol^{-1}; Q_3, −11.5 kcal mol^{-1}. Screening size of 100 million: minimum, −11.8 kcal mol^{-1}; maximum, −12.6 kcal mol^{-1}; median, −11.9 kcal mol^{-1}; Q_1, −11.8 kcal mol^{-1}; Q_3, −12.1 kcal mol^{-1}. Screening size of 1 billion: minimum, −12.3 kcal mol^{-1}; maximum, −13.4 kcal mol^{-1}; median, −12.4 kcal mol^{-1}; Q_1, −12.3 kcal mol^{-1}; Q_3, −12.6 kcal mol^{-1}.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The ready-to-dock library from Enamine is freely available online on the homepage of VirtualFlow at http://virtual-flow.org/real-library. Source Data for Figs. 2, 3 and Extended Data Figs. 7, 8 are available with the paper.

Code availability

VirtualFlow is mainly written in Bash (a Turing complete command language), which not only makes it simple for anyone to modify and extend the code, but also has essentially no computational overhead and is readily available in any major Linux distribution. The code for VirtualFlow is freely available on https://github.com/VirtualFlow, distributed under the GNU GPL open-source licence. The primary homepage for end users, which includes additional resources such as documentation, ligand libraries, tutorials and video demonstrations, is available at https://www.virtual-flow.org. The external docking programs discussed here are available as follows: AutoDock Vina is available at http://vina.scripps.edu, QuickVina 2 and QuickVina-W at https://qvina.github.io, Vina-Carb at http://glycam.org/docs/othertoolsservice/download-docs/publication-materials/vina-carb, Smina at https://sourceforge.net/projects/smina, AutoDockFR at http://adfr.scripps.edu and VinaXB at https://github.com/ssirimulla/vinaXB.

24. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform*. **3**, 33 (2011).
25. Morris, G. M. et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem*. **30**, 2785–2791 (2009).
26. Hutsell, S. Q., Kimple, R. J., Siderovski, D. P., Willard, F.S. & Kimple, A. J. High-affinity immobilization of proteins using biotin- and GST-based coupling strategies. *Methods Mol. Biol*. **627**, 75–90 (2010).
27. Hämäläinen, M. D. et al. Label-free primary screening and affinity ranking of fragment libraries using parallel analysis of protein panels. *J. Biomol. Screen*. **13**, 202–209 (2008).
28. Hulme, E. C. (ed.) *Receptor–Ligand Interactions: A Practical Approach* (Oxford Univ. Press, 1992).
29. Gans, P. et al. Stereospecific isotopic labeling of methyl groups for NMR spectroscopic studies of high-molecular-weight proteins. *Angew. Chem. Int. Ed*. **49**, 1958–1962 (2010).
30. Lu, M. et al. Discovery of a Keap1-dependent peptide PROTAC to knockdown Tau by ubiquitination-proteasome degradation pathway. *Eur. J. Med. Chem*. **146**, 251–259 (2018).
31. Irwin, J. J. et al. An aggregation advisor for ligand discovery. *J. Med. Chem*. **58**, 7076–7087 (2015).
32. LaPlante, S. R. et al. Compound aggregation in drug discovery: implementing a practical NMR assay for medicinal chemists. *J. Med. Chem*. **56**, 5142–5150 (2013).
33. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem*. **53**, 2719–2740 (2010).
34. Baell, J. B. & Nissink, J. W. M. Seven year itch: pan-assay interference compounds (PAINS) in 2017—utility and limitations. *ACS Chem. Biol*. **13**, 36–44 (2018).
35. Capuzzi, S. J., Muratov, E. N. & Tropsha, A. Phantom PAINS: problems with the utility of alerts for pan-assay interference compounds. *J. Chem. Inf. Model*. **57**, 417–427 (2017).

Acknowledgements We thank M. Zhang for help with the binding assays; the research computing teams of the Faculty of Arts and Sciences at Harvard University (especially S. Yockel, J. Cuff, F. Pontiggia and P. Edmon), the Jülich Supercomputing Centre, the Freie Universität (especially J. Dreger), the Harvard Medical School (HMS), the HLRN and the IT support of HMS (especially K. Bayer, G. Sekmokas and D. Morgan) for their support; K. E. Leigh, N. Gray, M. Kostic, A. Dubey, B. Klein, S. Schwaninger and S. Wu for discussions and manuscript preparation; the ICCB-Longwood Screening and East Quad NMR Facilities at HMS for assistance with the ligand screen; K. Arnett and the Center for Macromolecular Interactions at the HMS for advice on the SPR and BLI experiments; A. Jaffe for his support; and the teams from the Google Cloud Platform (especially S. Fang, R. Goldenbroit and D. Payne), Amazon Web Services, and Fluid Numerics for their support. This work was partially funded by a scholarship to C.G. from the Max Planck Institute for Molecular Genetics in Berlin and a scholarship from the Einstein Center for Mathematics Berlin. C.G. and K.F. thank the ECMath and MATHEON. C.G. is grateful to C. Schütte and P. Imhof for their support and supervision during his doctoral studies. We thank Z. Alirezaeizanjani, M. Bagherpoor and Anita Nivedha for testing VirtualFlow. M.H. acknowledges funding from Deutsche Forschungsgemeinschaft (CRC 958/Project A04, CRC 1114/Project A04). A.B. was supported by an Austrian Science Fund's Schrödinger Fellowship (J3872-B21) and an American Heart Association's fellowship (19POST34380800). This research was supported in part by grant TRT 0159 from the Templeton Religion Trust and by ARO Grant W911NF1910302 to A. Jaffe. K.M.P.D. was supported by a fellowship from the Max Kade Foundation and the Austrian Academy of Sciences. H.A. acknowledges funding from the Claudia Adams Barr Program for Innovative Cancer Research. G.W. acknowledges support from NIH grant CA200913, AI037581 and GM129026.

Author contributions C.G. conceived the project, and designed and implemented the drug discovery platform (VirtualFlow). H.A. and C.G. designed the experimental workflow. A.B., H.A., Z.-F.W., P.D.F. and K.M.P.D. designed and carried out the fluorescence polarization and NMR experiments. Z.-F.W. designed and carried out the SPR and BLI experiments. K.M.P.D. and Z.-F.W. carried out the dynamic light scattering experiments. M.H. provided technical assistance regarding the code and homepage. C.G. designed the applications (screening of KEAP1 and the preparation of the REAL library). P.W.C. analysed the NMR data. C.G. carried out the computations using VFLP (preparation of the REAL database) and VFVS (screening/rescoring of KEAP1). Y.S. Malets created the web interface to the REAL database. C.G. prepared the VirtualFlow homepage. Y.S. Moroz prepared the REAL database in the initial SMILES format. C.G. and Y.S. Moroz designed the structure of the VirtualFlow version of the REAL database. Y.S. Moroz, I.I. and D.S.R. supervised and directed the synthesis and purification of the on-demand compounds from the REAL library. D.A.S. helped to evaluate the screening hits. C.G., H.A., A.B., K.F., Z.-F.W., P.W.C. and G.W. prepared the manuscript. K.F., H.A. and G.W. supervised the project.

Competing interests I.I., D.S.R. and Y. S. Malets work for Enamine, a company that is involved in the synthesis and distribution of drug-like compounds. Y. S. Moroz is a scientific advisor for Enamine.

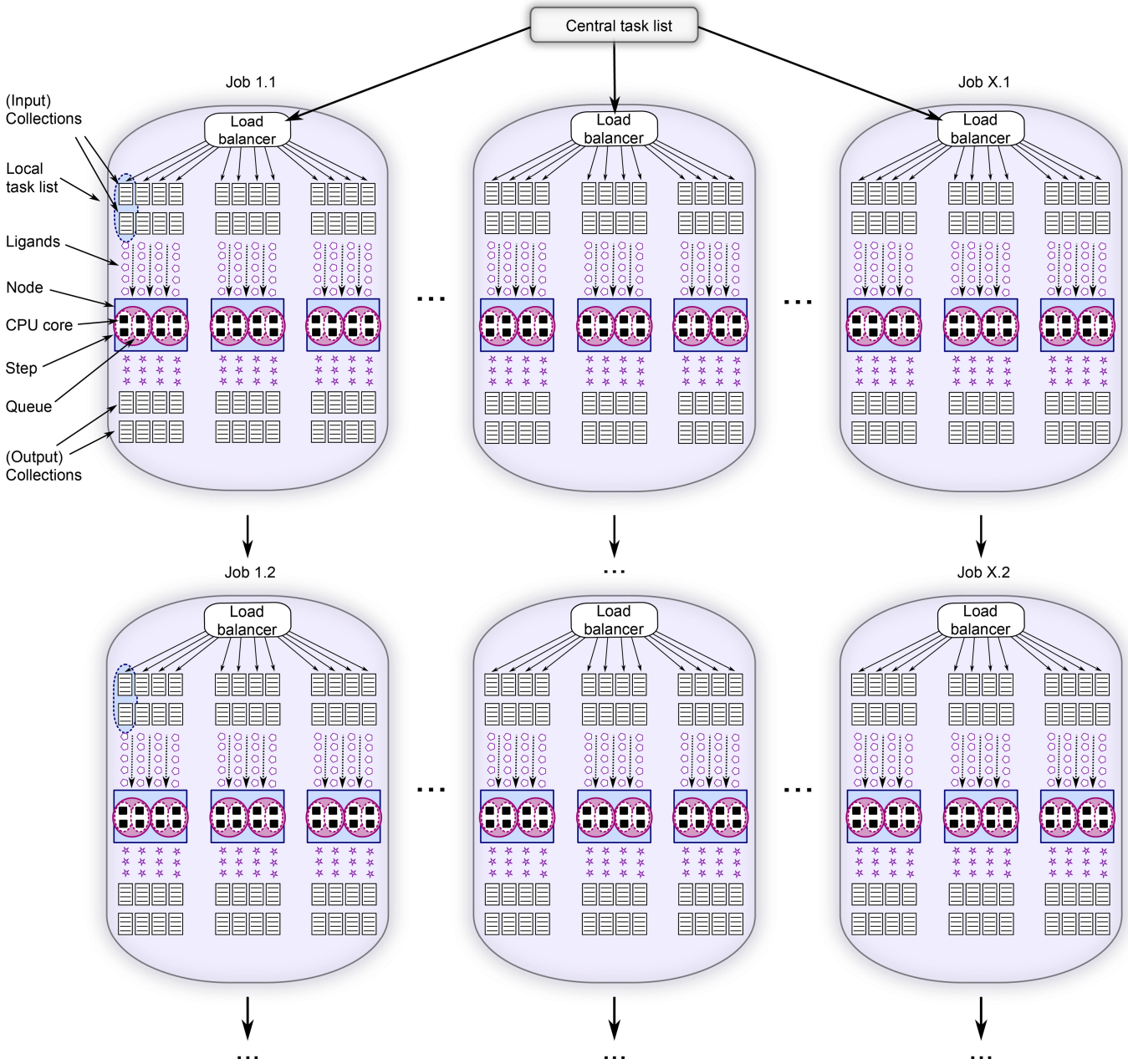
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2117-z.

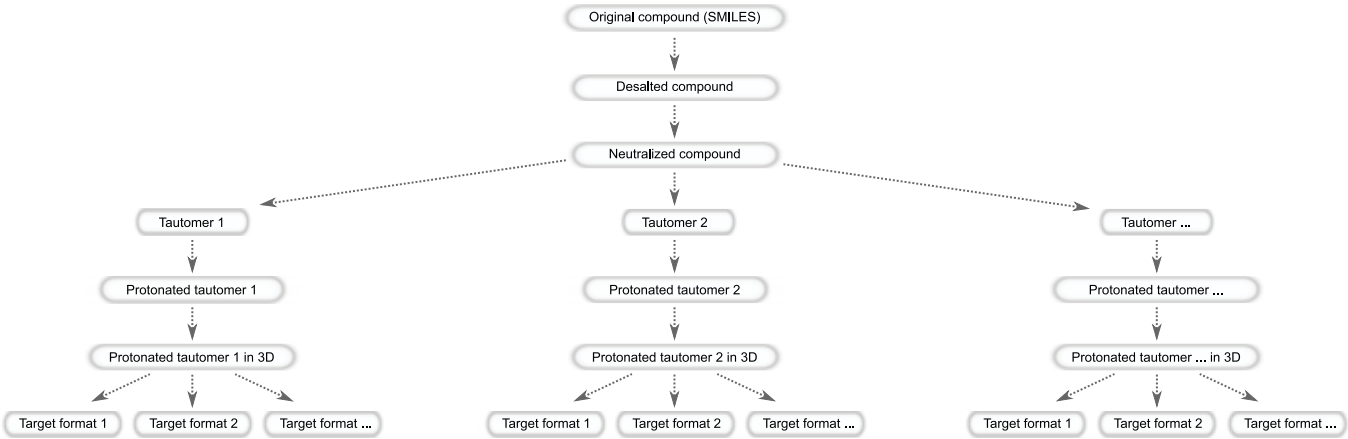
Correspondence and requests for materials should be addressed to C.G. or H.A.

Peer review information *Nature* thanks Tara Mirzadegan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

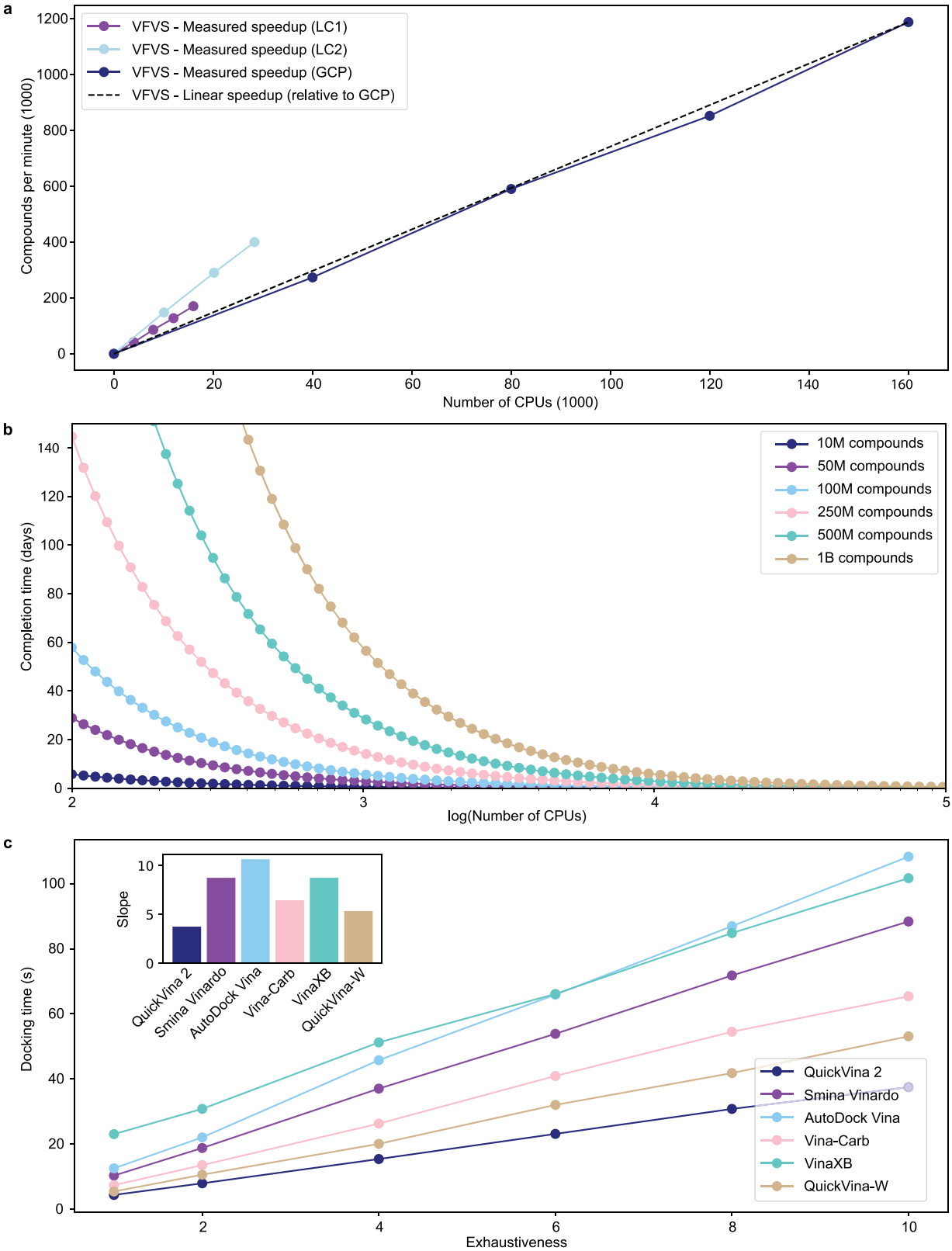
Reprints and permissions information is available at http://www.nature.com/reprints.



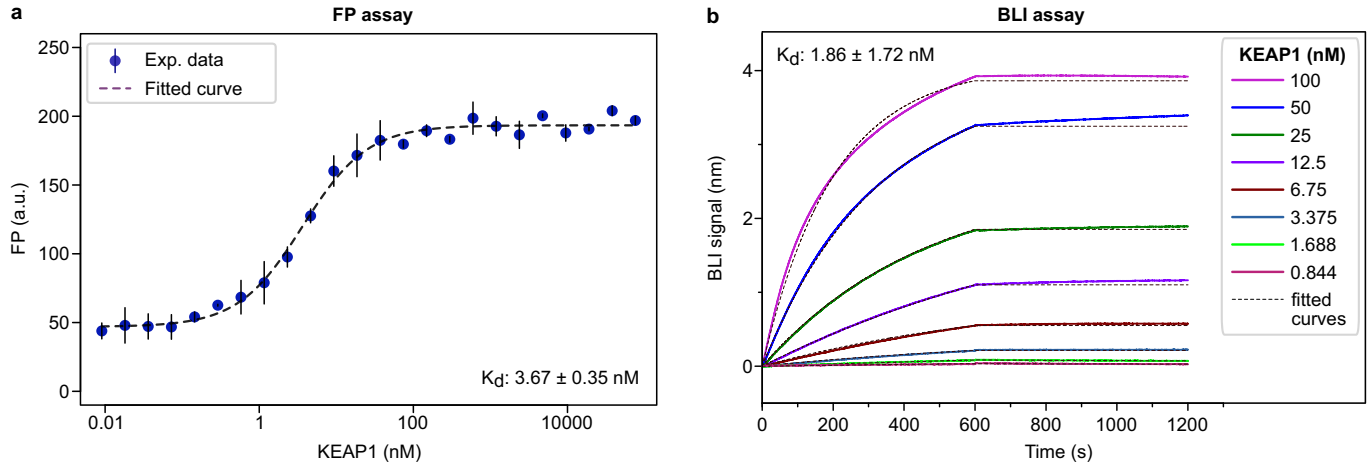
Extended Data Fig. 1 | Schematic overview of the organization of the VirtualFlow workflow on computer clusters. A computer cluster consists of compute nodes, that is, single computers (blue boxes), which contain a certain number of CPU cores (black squares inside the blue boxes). The resource manager (batch system) of the cluster generates so-called jobs (large violet ovals), each of which uses a certain number of CPU cores and nodes. In the example, each job uses three compute nodes, in which each node has eight CPU cores. Each job can contain multiple sub-jobs, referred to as job steps (purple circles). With VirtualFlow, each job step comprises multiple queues (white oval shapes within the purple circles). Often the workflow is set up such that on each CPU core one queue is running. Hierarchical multi-organization is required to allow VirtualFlow to run on any type of cluster, from the largest supercomputers (which often require that a single job has multiple nodes) to very small clusters (which often allow a job to use single CPU cores). Each queue processes ligands, which are taken from the input collections in raw form and stored in the output collection or database. The central task list contains all of the ligand collections that should be processed by the workflow, and they are distributed among the queues (into local task lists) by a workload balancer at the beginning of each job. The user can choose any number of batch system jobs (first row comprising job 1.1 to job X.1), which will automatically start successive jobs (second row comprising job 1.2 to job X.2) after their completion.



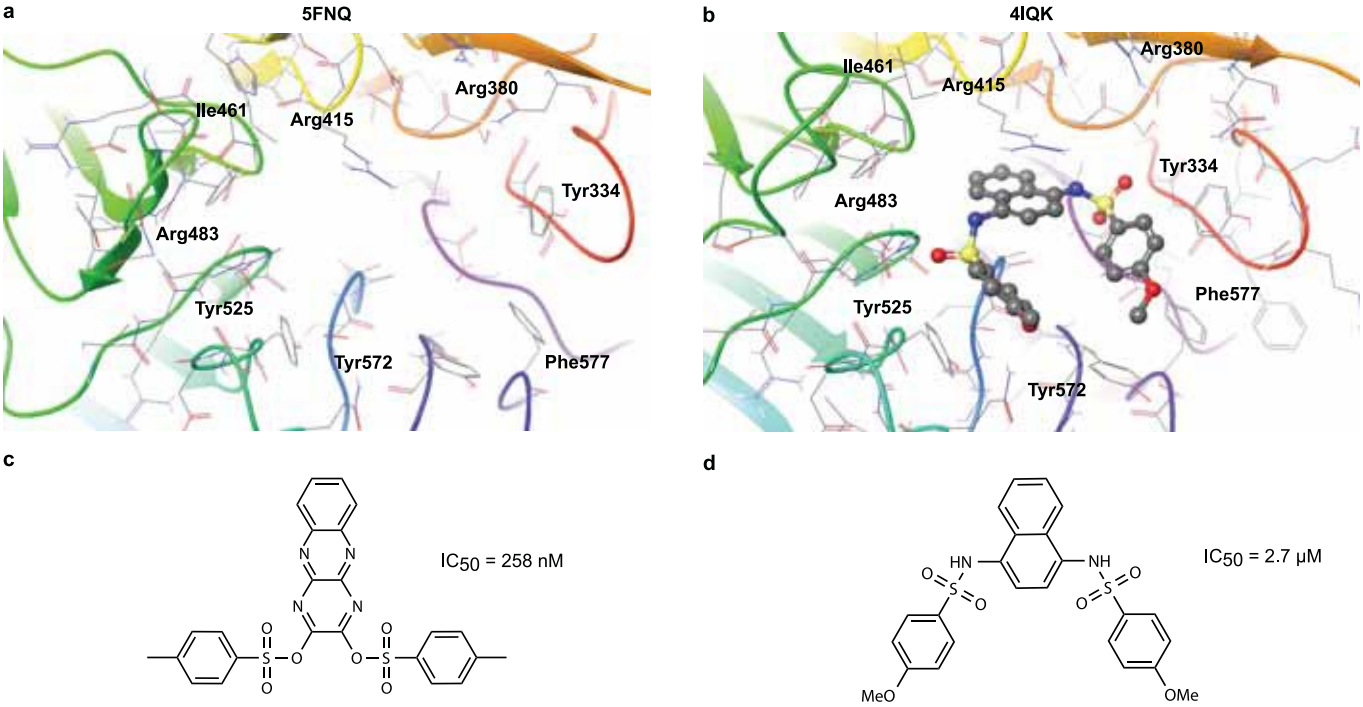
Extended Data Fig. 2 | Overview of possible processing steps during ligand preparation with VFLP. Ligands can be desalted, neutralized, and one, or possibly multiple, tautomeric state(s) as well as protonation states for each tautomer computed at specific pH values can be generated, three-dimensional coordinates can be computed and, finally, the molecules can be converted into one or potentially multiple desired target formats.



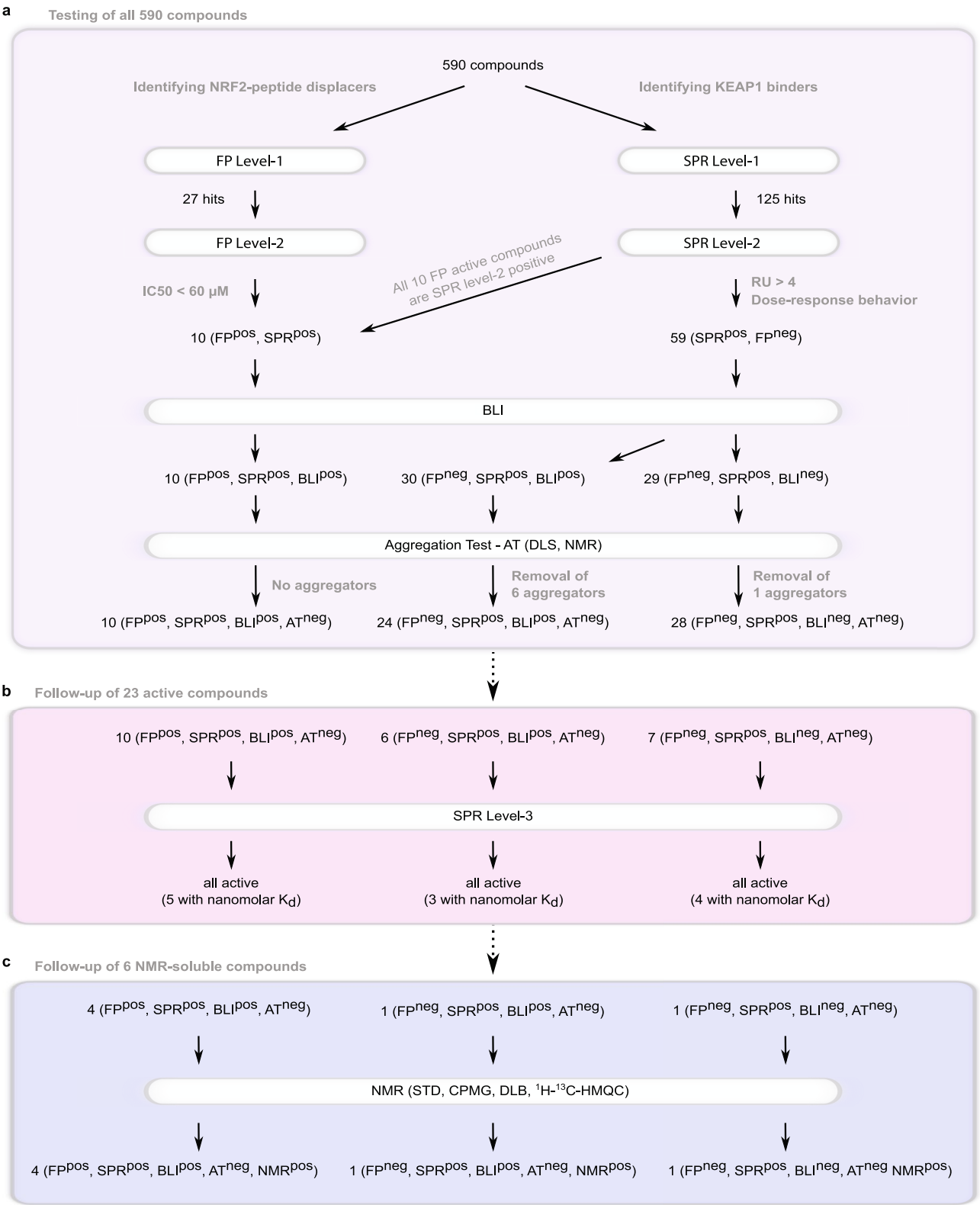
Extended Data Fig. 3 | Docking and virtual screening metrics. a, Scaling behaviour of VFVS using QuickVina 2 as the docking program. Tests with up to 30,000 cores on two local computer clusters (LC1 and LC2) and up to 160,000 CPUs on the GCP were carried out. The measured speedup is linear. DOVIS 2.0, an alternative software for virtual screenings on Linux computer clusters using AutoDock, was shown to exhibit near-linear scaling only up to 256 cores, as previously reported[4]. **b**, The computational time required (in days) for VFVS to complete virtual screens of different sizes, as a function of the number of CPUs being used in parallel. Each curve corresponds to a input ligand library with a different size, and the average computation time per ligand was assumed to be 5 s per ligand. **c**, Docking time of an average-sized ligand on a modern Intel CPU (using only a single core) as a function of the exhaustiveness parameter for different docking programs supported by VFVS. The bar plot in the inset shows the slope of the curves, which corresponds to the docking time per exhaustiveness unit. The test ligand that was used for this purpose is given by the SMILES code CN1CCN(S(=O)(=O)N2CCN(C(=O)CCCNC(=O)C3CC3)CC2)CC1. More detailed benchmarks can be found in publications related to these docking programs[5,12–17].



Extended Data Fig. 4 | Binding of the NRF2 peptide to KEAP1 as assayed by fluorescence polarization and BLI. a, a TAMRA-tagged NRF2 peptide was used for the fluorescence polarization (FP) assay. The fluorescence polarization assay was performed with three technical replicates per point. Data are mean ± s.d. for each titration point, along with the fitted curve. Two independent experiments were performed, each with similar results and one representative result is shown. **b**, A biotin-tagged NRF2 peptide was used for the BLI assay. The BLI experiment was repeated independently twicewith similar results and one representative result is shown.

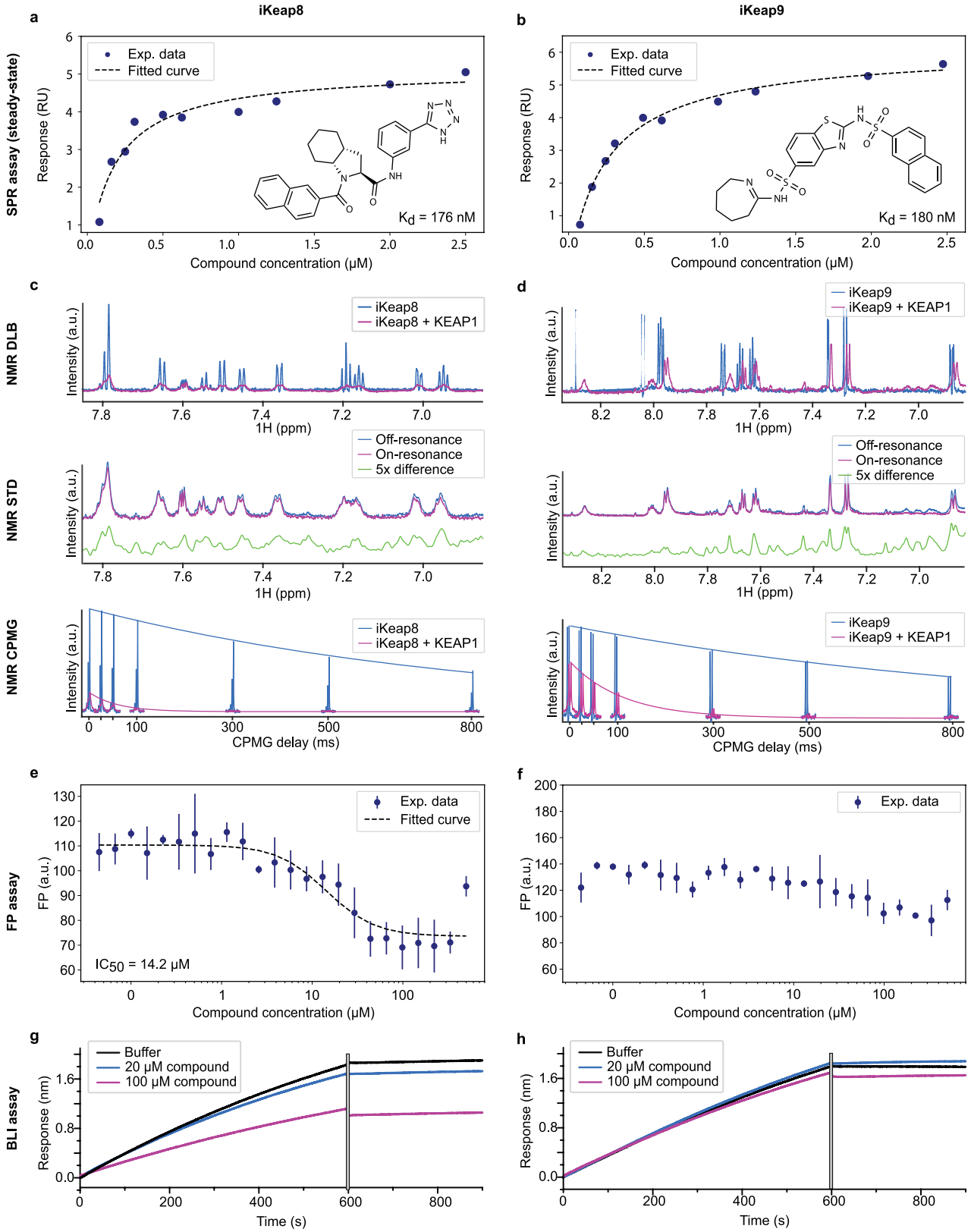


Extended Data Fig. 5 | Comparison of iKeap1 with the previously identified displacer C17. a, Crystal structure (PDB ID: 5FNQ)[9] of KEAP1 with its ligand removed, the structure used for the primary virtual screening procedure. **b**, Crystal structure of KEAP1 (PDB ID: 4IQK) with ligand C17 (Supplementary Table 1), the chemical structure of which is shown in **d**. **c**, **d**, iKeap1, the best displacer of the NFR2 peptide (**c**), is similar to compound C17, which has previously been identified by experimental methods (**d**). Although iKeap1 and C17 look similar, they differ in a number of aspects in their core scaffold (thus, analogues of the two compounds cover distinct chemical spaces, assuming that the analogues retain the core scaffold of the parent compound). This similarity, as well as the fact that the predicted docking positions (Fig. 3a) of both ligands (**b**) are nearly identical, is additional evidence that iKeap1 is binding at the predicted site.

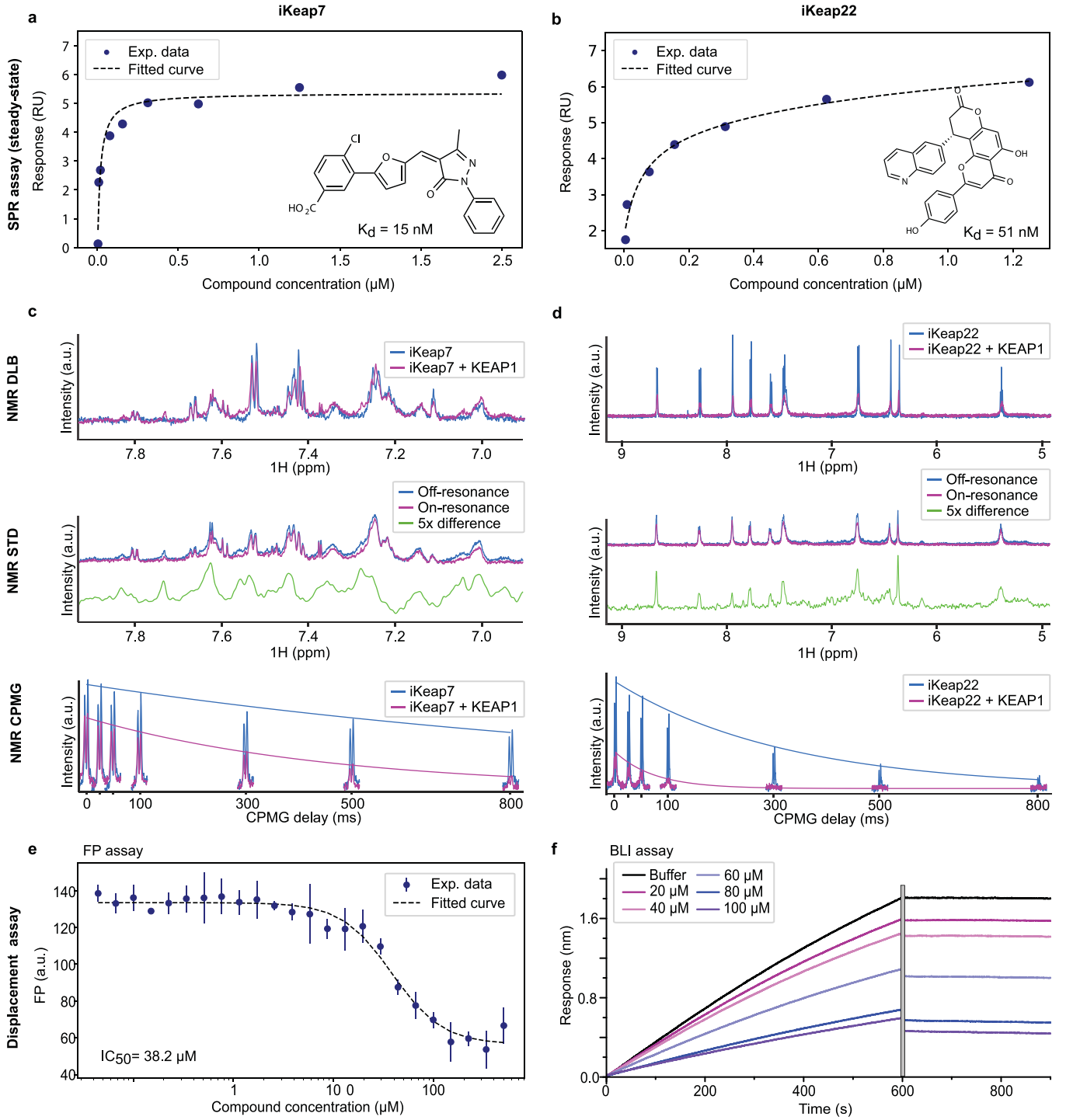


Extended Data Fig. 6 | See next page for caption.

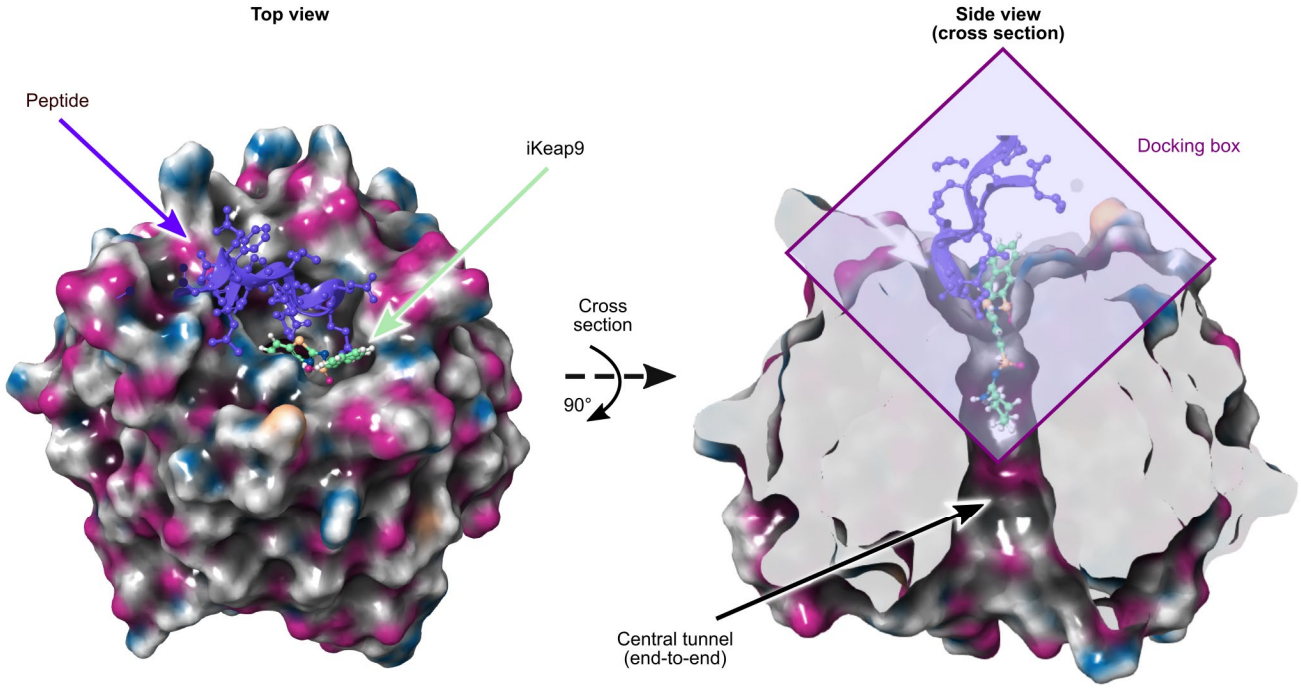
Extended Data Fig. 6 | Overview of binding assays to determine the activity of the hits identified by VirtualFlow. This schematic outlines the experimental validation workflow. The binding experiments can be broadly classified into two categories: (i) assays that directly detect the binding of the compounds to KEAP1 (SPR and NMR) and (ii) assays that detect the displacement of the NRF2 peptide from KEAP1 (fluorescence polarization and BLI). Compounds in level 2 SPR experiments were classified as active if they exhibited dose-dependent activity (measured over a range of five concentrations) and had an RU value greater than 4 at a compound concentration of 20 μM. **a**, The high-throughput workflow in which the 590 compounds identified as hits by VirtualFlow were tested using SPR and fluorescence polarization. The hits identified here were further validated by BLI and the potential of these hits to form aggregates was tested by DLS. **b**, Then, 23 of the potent hits were chosen for level 3 SPR analysis to measure accurate binding affinities. **c**, Six of the potent binders were further subjected to NMR analysis in both protein-detected and ligand-detected NMR experiments.



Extended Data Fig. 7 | Binder versus displacer. Here we highlight two scaffolds, iKeap8 and iKeap9, to illustrate the difference between binders and displacers. **a**, **b**, SPR confirms that both iKeap8 (**a**) and iKeap9 (**b**) bind to KEAP1 and with similar K_d values. Data are representative results from the SPR assay for iKeap8 and iKeap9. For each compound, three independent SPR experiments were performed, each with similar results and one representative result is shown. **c**, **d**, Ligand-detection NMR experiments shows that both iKeap8 (**c**) and iKeap9 (**d**) bind to KEAP1. **e**–**h**, However, fluorescence polarization (**e**, **f**) and BLI (**g**, **h**) assays show that iKeap8 (**e**, **g**) is able to displace the NRF2 peptide whereas iKeap9 (**f**, **h**) is not able to effectively displace the NRF2 peptide. The fluorescence polarization assay was performed with three technical replicates per concentration measured. Data are mean ± s.d. for each titration point shown together with the fitted curve.



Extended Data Fig. 8 | Displacers validated by fluorescence polarization and BLI. Here we show two more displacers, iKeap7 and iKeap22. **a**, **b**, Both iKeap7 (**a**) and iKeap22 (**b**) were confirmed as binders by SPR. **c**, **d**, Ligand-detection NMR experiments show that both iKeap7 (**c**) and iKeap22 (**d**) bind to KEAP1. **e**, iKeap7 is confirmed to be a displacer of the NRF2 peptide by both fluorescence polarization and BLI (data not shown). **f**, As the fluorescence polarization experiments of iKeap22 were affected by autofluorescence, BLI was needed to confirm that this compounds is a displacer. The fluorescence polarization assay was performed with three technical replicates per concentration measured. Data are mean ± s.d. for each titration point, shown along with the fitted curve. Two independent BLI experiments were performed with similar results and one representative result is shown here.



Extended Data Fig. 9 | NRF2 peptide- and ligand-binding sites, rationale for binder versus displacer. Here we show the docking pose of one of the hit compounds (iKeap9, green ball-and-stick representation) bound to KEAP1, together with the NRF2 peptide (PDB ID: 4IFL; peptide in violet). iKeap9 is a tight binder (180 nM by steady-state SPR) but cannot displace NRF2. Left, the top view. Right, the side view of the cross-section of KEAP1 along the central plane. The violet box indicates the docking region (where the ligands were allowed to bind), which was used in the virtual screening. The site of interest includes a part of the deep pocket/tunnel of the β-barrel-shaped KEAP1, as it enables ligands to bind more tightly by insertion into the channel than on a shallow surface. However, the deep tunnel is largely non-overlapping with the peptide-binding site (which binds to the entrance site of the tunnel). Thus, binding molecules might only partially interfere with peptide binding, which could reduce or eliminate the ability of small-molecule binders to displace the peptide. The ability of a small molecule to displace the peptide is difficult to predict, and was not attempted in this study. In some cases, small molecules can also act as molecular glues and strengthen the interaction between NRF2 and KEAP1.

nature research

Corresponding author(s): Gerhard Wagner, Haribabu Arthanari, Christoph Gorgulla

Last updated by author(s): Feb 18, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Software Autodock Vina, Smina, QuickVina and Topspin (Bruker) used in this study are all publicly available. The platform VirtualFlow which was developed for this manuscript is freely available on GitHub the following URL https://github.com/VirtualFlow Additional resources inducing documentation, ligand libraries, tutorials and video demonstration are freely available at the VirtualFlow homepage http://virtual-flow.org/
Data analysis	DataWarrior, Python, Maestro, Matlab, Matlab, R, TopSpin (Bruker); (all publicly available)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The ready-to-dock library from Enamine is available online on the homepage of VirtualFlow at http://virtual-flow.org/real-library. Other datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In experiments where we performed replicated the sample size and the nature of the replicated (technical and/or independent) is clearly mentioned. Sample size is not applicable to some of experimental measurements (NMR) described in this manuscript. The manuscript contains no animal data.
Data exclusions	No data was excluded in this study.
Replication	We had performed replicates (independent and/or technical replicates) for the experiments that were used to extract quantitative information, such as Kd (SPR) and IC50 (FP) and DLS. We have updated the figure captions and the methods to reflect this and provided explicit details of how the replicates were done. We have also provided the source data where applicable. Data replication does not apply to NMR measurements described in this manuscript.
Randomization	Randomization does not apply to this study since we are not comparing data across samples or cohorts.
Blinding	Blinding is not relevant since the study does not compare different groups of measurements

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function

https://doi.org/10.1038/s41586-020-2195-y

Received: 9 December 2019

Accepted: 18 March 2020

Published online: 22 April 2020

Check for updates

Svetlana O. Dodonova[1], Fangjie Zhu[2], Christian Dienemann[1], Jussi Taipale[2] & Patrick Cramer[1]✉

'Pioneer' transcription factors are required for stem-cell pluripotency, cell differentiation and cell reprogramming[1,2]. Pioneer factors can bind nucleosomal DNA to enable gene expression from regions of the genome with closed chromatin. SOX2 is a prominent pioneer factor that is essential for pluripotency and self-renewal of embryonic stem cells[3]. Here we report cryo-electron microscopy structures of the DNA-binding domains of SOX2 and its close homologue SOX11 bound to nucleosomes. The structures show that SOX factors can bind and locally distort DNA at superhelical location 2. The factors also facilitate detachment of terminal nucleosomal DNA from the histone octamer, which increases DNA accessibility. SOX-factor binding to the nucleosome can also lead to a repositioning of the N-terminal tail of histone H4 that includes residue lysine 16. We speculate that this repositioning is incompatible with higher-order nucleosome stacking, which involves contacts of the H4 tail with a neighbouring nucleosome. Our results indicate that pioneer transcription factors can use binding energy to initiate chromatin opening, and thereby facilitate nucleosome remodelling and subsequent transcription.

Transcription of the human genome is controlled by about 1,600 transcription factors[4]. Transcription factors recognize DNA motifs and recruit protein complexes that enable transcription initiation[5]. The binding of most transcription factors is restricted to regions of the genome that are not packaged into chromatin[6]. Some transcription factors can, however, bind to chromatin via contacts to its fundamental unit, the nucleosome[7]. These pioneer transcription factors can initiate transcription in silent chromatin regions[8], and are required for embryo development, cell differentiation and cell reprogramming[9].

SOX2 and OCT4 (also known as POU5F1) are pioneer factors that are widely used for the reprogramming of adult cells to induced pluripotent stem cells[2,10,11]. SOX2 and OCT4 can interact with nucleosomes in vitro and in vivo[12,13]. SOX2 alone can direct chromatin opening[14] and bind target DNA sites before OCT4 in vivo[10], and SOX2 binding to DNA can also follow OCT4 binding in vitro[15]. Most factors in the SOX family (hereafter, SOX factors) show pioneer factor function[7] and are essential for developmental processes[1,9]. The mutation of SOX factors can lead to severe developmental defects and cancer[16]. How pioneer transcription factors such as SOX factors bind to the nucleosome, and how they make DNA accessible for their non-pioneer partner proteins, is unknown.

To investigate this, we determined the cryo-electron microscopy (cryo-EM) structure of human SOX2 in complex with a nucleosome (Methods). We used a 147-bp nucleosomal DNA sequence (hereafter referred to as DNA-1) (Extended Data Fig. 1) that was previously selected for binding the closely related factor SOX11[7]. The DNA-binding domains (DBDs) of SOX2 and SOX11 share 83% sequence similarity (Extended Data Fig. 2), and bind the same DNA motif (TTGT)[17]. Base pairs of the TTGT core motif are specifically contacted by amino acid residues in the RPMNAFMVW signature motif of the SOX-factor HMG box[18]. SOX2 and SOX11 bind the same target sites in cells, substantially differ only in regions that flank their DBDs, and recruit different factors[19,20].

Nucleosomes containing DNA-1 bound recombinant SOX2 or SOX11 DBDs (Extended Data Fig. 3). We added the purified SOX2 DBD in excess to reconstituted nucleosomes, plunge-froze cryo-EM grids and collected cryo-EM data (Methods). A subset of 32,301 particles resulted in an approximately 5.5 Å resolution map that showed extra density on the nucleosome surface (Fig. 1, Extended Data Fig. 4, Extended Data Table 1). We fitted the map with structures of the nucleosome[21] and the SOX2 DBD[22] (Extended Data Figs. 4, 5).

The nucleosome–SOX2 structure that we obtained revealed a single copy of the SOX2 DBD bound to DNA at superhelical location (SHL) +2 (Fig. 1). The observed SOX binding involves specific interactions with the DNA motif, as confirmed by site-directed mutagenesis of involved residues in SOX11 and by mutagenesis of the DNA-1 sequence (Extended Data Figs. 3, 6). In further agreement with our structure, in vivo[23] SOX factors preferentially occupy target sites that are located near the centre of the nucleosome[24]. Although DNA-1 contains multiple SOX-binding motifs and can bind multiple copies of SOX2, nucleosomes containing DNA-1 bind only one copy of SOX2 (Extended Data Figs. 1, 3). In the context of the DNA-1 sequence, binding of the SOX DBD to other sites on the nucleosome would result in clashes with DNA and histones (Extended Data Fig. 1d).

Despite extensive efforts, the resolution of our nucleosome–SOX2 structure remained limited. We therefore determined the structure of a nucleosome bound to the DBD of SOX11 (Methods). The cryo-EM dataset contained 222,731 particles and resulted in a detailed reconstruction

[1]Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. [2]Department of Biochemistry, University of Cambridge, Cambridge, UK. ✉e-mail: patrick.cramer@mpibpc.mpg.de

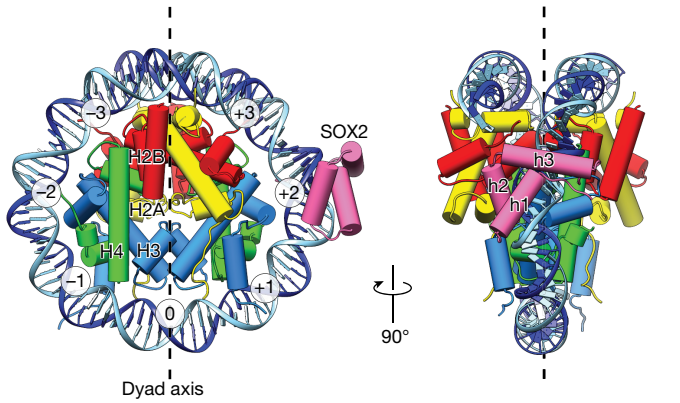


Fig. 1 | Structure of the nucleosome–SOX2 complex. Structure of the nucleosome–SOX2 complex reveals SOX2 binding at SHL +2. Top and side views are related by a 90° rotation around the dyad (dashed line). Superhelical locations −3 to +3 are labelled. SOX2 DBD is shown in pink; histones H2A, H2B, H3 and H4 are shown in yellow, red, blue and green, respectively; DNA is shown in dark and light blue. h1–h3, helices 1–3.

at 3.7 Å resolution (Extended Data Figs. 4, 5). To aid model building, we additionally determined the crystal structure of the SOX11 DBD in complex with a DNA fragment at 2.5 Å resolution (Extended Data Fig. 2c–f, Extended Data Table 2).

The structure of the nucleosome–SOX11 complex was virtually identical to that of the nucleosome–SOX2 complex (Extended Data Figs. 4g, 5). It is also a good model for nucleosome complexes with the DBDs of other members of the SOX family, which are highly conserved (Extended Data Fig. 2). For comparisons, we further determined the structure of the free nucleosome containing DNA-1 from 368,270 particles at 3.2 Å resolution (Extended Data Figs. 4, 5). This structure was highly similar to the canonical nucleosome structure (Protein Data Bank (PDB) code 6FQ5), root mean square deviation (r.m.s.d.) (P) = 1.0 Å).

Comparison of the nucleosome–SOX11 structure with the free nucleosome structure shows that SOX11 binding leads to strong local DNA distortions at SHL +2 (Fig. 2) (local r.m.s.d. (P) = 3.9 Å; calculated for 12 bp of DNA). SOX11 widens the DNA minor groove by 7 Å and pulls the DNA away from the histone octamer by 3–4 Å (coordinate error of approximately 1 Å), which increases DNA bending (Fig. 2). These DNA distortions are induced by SOX11 binding, and are also observed in our SOX11–DNA crystal structure (r.m.s.d. (P) = 1.4 Å, for 12 bp of DNA). Thus, the SOX factor uses binding energy to distort DNA locally, despite competing histone–DNA interactions.

In both nucleosome–SOX-factor structures, approximately 2.5 turns of both DNA termini are detached from the histone octamer and not visible in the cryo-EM densities (Fig. 3, Extended Data Fig. 5). This is consistent with the observation that several SOX factors facilitate DNA unwrapping from the nucleosome[7], and with the known high dynamics of the terminal DNA[25,26]. A DNA cleavage assay supports the increase in accessibility of the terminal nucleosomal DNA in the presence of SOX11 (Extended Data Fig. 7). Comparison with the free nucleosome structure indicates that terminal DNA at SHL −7, SHL −6 and SHL −5 is detached from the octamer because of a clash with helix 2 of the SOX factor (Fig. 3c, Supplementary Video 1). Thus SOX factor binding to the nucleosome facilitates DNA detachment and increases accessibility of terminal DNA.

Our cryo-EM data also suggest the dynamics that underlie nucleosome invasion by SOX factors. A set of particles from a separate dataset (151-bp DNA-1) (Methods) resulted in an alternative nucleosome–SOX11 structure in which the terminal DNA near SOX11 remained associated with the histone octamer (hereafter, nucleosome–SOX11*) (Fig. 3, Extended Data Figs. 4, 5). Thus, SOX factors may initially bind to their target site without detaching the second DNA gyre. Movement of the DNA-bound SOX factor to the position observed in the nucleosome–SOX11 structure would then lead to a clash that is resolved by terminal DNA detachment. This resulted in a model of nucleosome invasion and DNA unwrapping by SOX-factor binding (Fig. 3, Supplementary Video 1). The proposed mechanism differs from that used by the yeast pioneer factor Reb1, which binds and traps terminal DNA[27].

In our nucleosome–SOX factor structures, terminal DNA is detached on both sides of the nucleosome, which suggests an additional allosteric effect of the SOX factor on nucleosome stability. Detachment of terminal DNA on the other side of the nucleosome may be stabilized by binding of a second copy of the SOX factor at SHL −2, which we observed in a subpopulation of our cryo-EM particles (Extended Data Figs. 4, 5e). In this nucleosome–$SOX11_2$ structure, the orientation of SOX-factor binding is determined by the asymmetric DNA motifs at both SHL +2 and SHL −2, with the latter apparently having lower affinity (Extended Data Figs. 3, 6). We speculate that SOX factors may also bind to multiple sites of a nucleosome in vivo. For example, a well-studied SOX2-binding genomic location (LIN28) contains two canonical SOX2 DNA motifs within a nucleosome and shows a broad peak of SOX2 occupancy[12].

The nucleosome–$SOX11_2$ structure shows that SOX11 binding at SHL −2 is incompatible with binding of terminal DNA at SHL +7, SHL +6 and SHL +5, although the predicted clash at this location is with helix 3 and both termini of the SOX-factor DBD. DNA detachment is also observed with the use of Förster resonance energy transfer experiments when SOX11 is present at high concentrations (Extended Data Fig. 7). Thus, SOX factors can induce detachment of both DNA ends and can bind to both sides of the nucleosome (Fig. 3, Supplementary Video 1). These observations agree with the recently described strong preference for SOX2 binding at approximately ±25 bp around

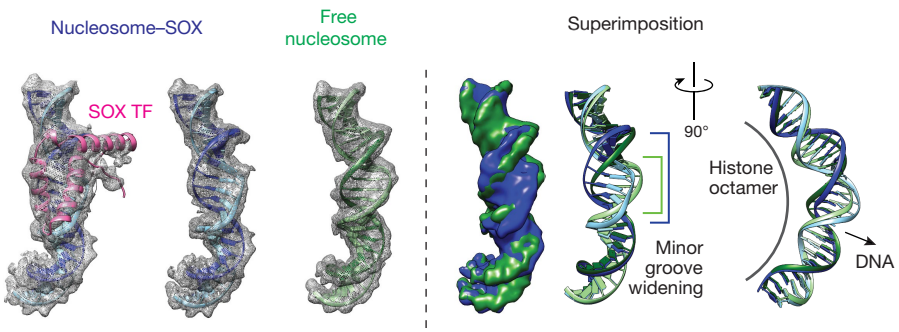


Fig. 2 | Structure of the nucleosome–SOX11 complex, and local DNA distortion. SOX11 is shown in pink, and DNA bound by SOX is shown in dark and light blue. DNA in the free nucleosome structure is shown in dark and light green. The cryo-EM maps shown here were Gaussian-smoothened. For clarity, SOX density was segmented out on the right (blue models). TF, transcription factor.

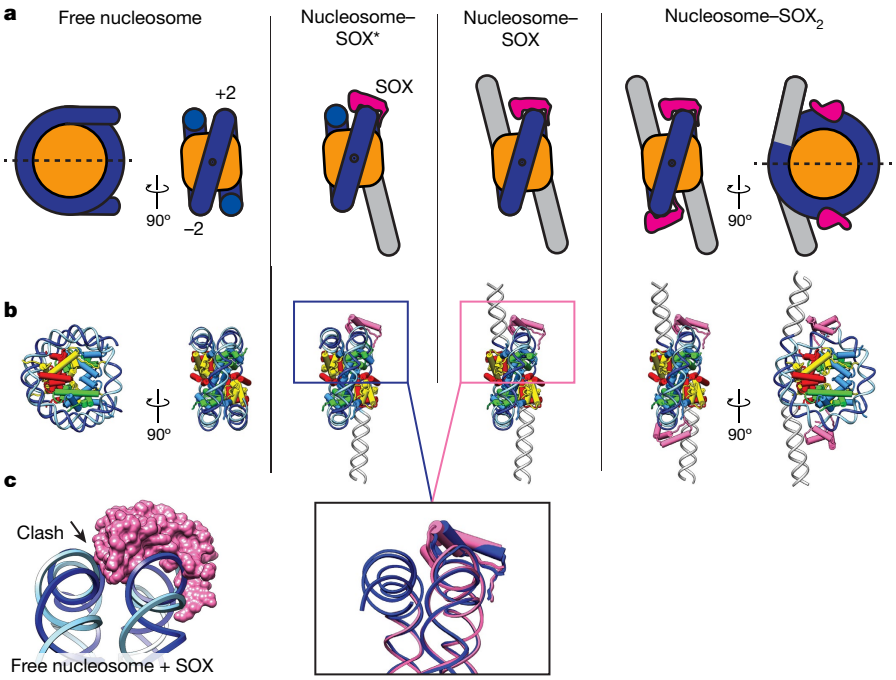


Fig. 3 | Model of nucleosome invasion by SOX factors. a, Nucleosome invasion by SOX factors and terminal DNA detachment. Schematic of the structures reported here. From left to right, free nucleosome, nucleosome–SOX11*, nucleosome–SOX11 and nucleosome–SOX11$_2$ are shown. The histone octamer is shown in orange, SOX in pink and DNA in blue. Detached DNA is shown in grey. The dyad is shown as a dashed line or as a dot. **b**, Four structures, coloured as in Fig. 1. Detached DNA was modelled as ideal B-DNA (grey). The black box shows a comparison of the nucleosome–SOX* (dark blue) and nucleosome–SOX (pink) structures. **c**, DNA superposition in the free nucleosome and the SOX factor (surface view) from the nucleosome–SOX structure illustrates the clash between SOX and the second DNA gyre.

the nucleosome dyad in vivo[24]. However, the possibility that SOX factors may also bind additional nucleosomal positions in other contexts is not excluded.

The nucleosome–SOX11 structure further shows that binding of SOX11 repositions the N-terminal tail of H4 (Fig. 4, Extended Data Fig. 8). In the free nucleosome structure, the H4 tail binds to its canonical site and follows a trajectory towards DNA at SHL +2. However, in the nucleosome–SOX11 structure, the binding site of the H4 tail at SHL +2 is occupied by the SOX11 C-terminal tail (Fig. 4, Extended Data Fig. 8). The H4 tail is displaced, rotated by about 90° and extends towards SHL +1. The functionally important residue lysine 16 (K16) moves by around 33 Å. However, at SHL −2 SOX11 is oriented differently and does not displace the H4 tail (Extended Data Fig. 8c).

We speculate that SOX-factor binding may be incompatible with the formation of canonical nucleosome–nucleosome contacts[28] (Extended Data Fig. 8). Formation of nucleosome arrays depends on the H4 tail and is impaired by K16 acetylation or tail truncation[29–31]. Nucleosome stacking is mediated by H4 tail residues K16–R19 that interact with the acidic patch of the H2A–H2B histone dimer of the neighbouring nucleosome[21,32]. Modelling the SOX DBD onto a nucleosome array[32] suggests that the pioneer factor could be accommodated. SOX binding at SHL +2 and SHL −2 might be preferred over binding at the nucleosome dyad, which would be occluded by the H1 linker histone. However, for efficient chromatin opening, SOX factors cooperate with other transcription factors such as OCT4, KLF4, PAX6, Nanog, BRN2, and PRX1[16], and with ATP-consuming chromatin remodellers[33].

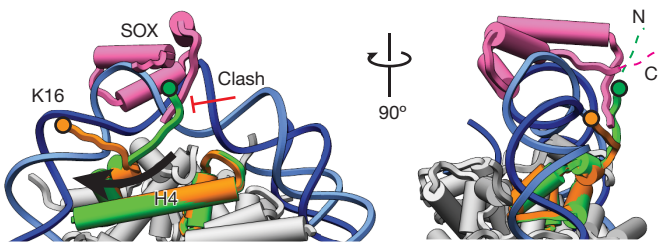


Fig. 4 | SOX11 repositions the H4 tail. Top and side views of SHL +2 with SOX transcription factor (pink). Histones are grey, except for H4. H4 from the free nucleosome is shown in green where the H4 N-terminal (N) tail would clash with the C-terminal (C) SOX tail. In the nucleosome–SOX structure, the H4 tail is repositioned (orange). Residue K16 is marked with a coloured circle.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2195-y.

1. Iwafuchi-Doi, M. & Zaret, K. S. Cell fate control by pioneer transcription factors. *Development* **143**, 1833–1837 (2016).
2. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
3. Adachi, K., Suemori, H., Yasuda, S. Y., Nakatsuji, N. & Kawase, E. Role of SOX2 in maintaining pluripotency of human embryonic stem cells. *Genes Cells* **15**, 455–470 (2010).
4. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
5. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription *in vivo*. *Nature* **461**, 186–192 (2009).
6. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
7. Zhu, F. et al. The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76–81 (2018).
8. Cirillo, L. A. et al. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell* **9**, 279–289 (2002).
9. Boller, S., Li, R. & Grosschedl, R. Defining, B cell chromatin: lessons from EBF1. *Trends Genet.* **34**, 257–269 (2018).
10. Chen, J. et al. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell* **156**, 1274–1285 (2014).

11. Velychko, S. et al. Excluding Oct4 from Yamanaka cocktail unleashes the developmental potential of iPSCs. *Cell Stem Cell* **25**, 737–753.e4 (2019).
12. Soufi, A. et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).
13. Meers, M. P., Janssens, D. H. & Henikoff, S. Pioneer factor–nucleosome binding events during differentiation are motif encoded. *Mol. Cell* **75**, 562–575.e5 (2019).
14. Malik, V. et al. Pluripotency reprogramming by competent and incompetent POU factors uncovers temporal dependency for Oct4 and Sox2. *Nat. Commun.* **10**, 3477 (2019).
15. Biddle, J. W., Nguyen, M. & Gunawardena, J. Negative reciprocity, not ordered assembly, underlies the interaction of Sox2 and Oct4 on DNA. *eLife* **8**, e41017 (2019).
16. Kamachi, Y. & Kondoh, H. Sox proteins: regulators of cell fate specification and differentiation. *Development* **140**, 4129–4144 (2013).
17. Badis, G. et al. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
18. Jauch, R., Ng, C. K., Narasimhan, K. & Kolatkar, P. R. The crystal structure of the Sox4 HMG domain–DNA complex suggests a mechanism for positional interdependence in DNA recognition. *Biochem. J.* **443**, 39–47 (2012).
19. Bergsland, M. et al. Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.* **25**, 2453–2464 (2011).
20. Wiebe, M. S., Nowling, T. K. & Rizzino, A. Identification of novel domains within Sox-2 and Sox-11 involved in autoinhibition of DNA binding and partnership specificity. *J. Biol. Chem.* **278**, 17901–17911 (2003).
21. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–260 (1997).
22. Williams, D. C., Jr, Cai, M. & Clore, G. M. Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2. Hoxb1–DNA ternary transcription factor complex. *J. Biol. Chem.* **279**, 1449–1457 (2004).
23. Voong, L. N. et al. Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell* **167**, 1555–1570.e15 (2016).
24. Li, S., Zheng, E. B., Zhao, L. & Liu, S. Nonreciprocal and conditional cooperativity directs the pioneer activity of pluripotency transcription factors. *Cell Rep.* **28**, 2689–2703 (2019).
25. Hall, M. A. et al. High-resolution dynamic mapping of histone–DNA interactions in a nucleosome. *Nat. Struct. Mol. Biol.* **16**, 124–129 (2009).
26. Bilokapic, S., Strauss, M. & Halic, M. Histone octamer rearranges to adapt to DNA unwrapping. *Nat. Struct. Mol. Biol.* **25**, 101–108 (2018).
27. Donovan, B. T., Chen, H., Jipa, C., Bai, L. & Poirier, M. G. Dissociation rate compensation mechanism for budding yeast pioneer transcription factors. *eLife* **8**, e43008 (2019).
28. Pepenella, S., Murphy, K. J. & Hayes, J. J. Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. *Chromosoma* **123**, 3–13 (2014).
29. Gordon, F., Luger, K. & Hansen, J. C. The core histone N-terminal tail domains function independently and additively during salt-dependent oligomerization of nucleosomal arrays. *J. Biol. Chem.* **280**, 33701–33706 (2005).
30. Shogren-Knaak, M. et al. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* **311**, 844–847 (2006).
31. Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *J. Mol. Biol.* **327**, 85–96 (2003).
32. Song, F. et al. Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science* **344**, 376–380 (2014).
33. Engelen, E. et al. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. *Nat. Genet.* **43**, 607–611 (2011).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Plasmids and strains

Full-length histone sequences from *Homo sapiens* were incorporated into the following plasmids: pET22B-H2B, pET22b-H3.2, pET3a-H4 (kindly provided by the W. Fischle laboratory). The H2A construct was cloned into a LIC1B vector (MacroLabs) and contained an N-terminal 6×His-tag followed by a tobacco etch virus (TEV) protease cleavage site (HHHHHHENLYFQS). The SOX2 DBD construct contained residues 36–121 of the full-length SOX2 (UniProt ID P48431). The DBD sequence was codon-optimized and synthesized by IDT as a gBlock. The gBlock was inserted into a LIC1B plasmid following N-terminal 6×His-tag and a TEV protease cleavage site sequences. The SOX11 DBD with short flanking sequences contained residues 33–138 of full-length SOX11 (UniProt ID P35716). It was inserted into a LIC1B plasmid. The construct was identical to the SOX11 construct used in a previous study[7]. Protein constructs are schematically shown in Extended Data Fig. 2.

Protein purification

Histones were purified according to standard protocols[34,35]. Purified histones were flash-frozen and lyophilized. Histones were resuspended in unfolding buffer (6 M guanidine hydrochloride, 20 mM HEPES pH 7.5, 10 mM dithiothreitol (DTT)). H2A, H2B, H3 and H4 were mixed in 1.2:1.2:1:1 ratios, and dialysed against three changes of refolding buffer high (RB high: 20 mM HEPES pH 7.5, 1 mM EDTA, 2 M NaCl, 2 mM DTT). After dialysis, the sample was concentrated and loaded onto a size-exclusion chromatography column (Superdex 200 10/300 GL, GE Healthcare). A peak corresponding to the complete octamer was collected and used for nucleosome reconstitution. The SOX2 DBD was expressed in BL21 (DE3) RIL *Escherichia coli* cells and purified over a sequence of columns: affinity His-Trap HP, cation exchange HiTrap SP-HP and size-exclusion Superdex75 10/300 GL (GE Healthcare). The His-tag was cleaved off after the affinity purification step. The SOX11 DBD was expressed and purified exactly as the SOX2 DBD. Purified proteins in the final buffer (20 mM HEPES pH 7.5, 1 mM EDTA, 150 mM NaCl, 1 mM DTT) were flash-frozen and stored at −80°.

DNA preparation

DNA-1 template sequence was: ATCTACACGACGCTCTTCCGATCTAATTT ATGTTTGTTAGCGTTATACTATTCTAATTCTTTGTTTCGGTGGTATTGTTT ATTTTGTTCCTTTGTGCGTTCAGCTTAATGCCTAACGACACTCGGAGAT CGGAAGAGCACACGTGAT. This sequence was directly (no changes) adopted from the NCAP-SELEX experiment with nucleosomes and SOX transcription factor described previously[7].

DNA-1a sequence with all but one of the TTGT motifs replaced by a random sequence was: ATCTACACGACGCTCTTCCGATCTAATTTATTCAG ACTAGCGTTATACTATTCTAATTTCAGACTTCGGTGGTCAGACTTATCAG ACTCCTTTGTGCGTTCAGCTTAATGCCTAACGACACTCGGAGATCGG AAGAGCACACGTGAT.

Widom 601 DNA template used as a control was: ATCGAGAATCCC GGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTA AACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTA CTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCCGAT. Three bases at each end were changed to accommodate an EcoRV restriction site.

The DNA* template used for the nucleosome–SOX* structure determination was 151 bp long and almost identical to the DNA-1: A TC**CC**TACACGACGCTCTTCCGATCTAATTTATGTTTGTTAGCGTTATAC TATTCTAATTCTTTGTTTCGGTGGTATTGTTTATTTTGTTCCTTTGTGC GTTCAGCTTAATGCCTAACGACACTCGGAGATCGGAAGAGCACACGT**C T**GAT. Two additional nucleotides on each side in the DNA* template are highlighted in bold. The rational for using a slightly longer 151-bp DNA construct was the following. The H2A C-terminal tail regulates nucleosome conformation by binding to linker DNA at different locations and stabilizes the nucleosome[36]. When a longer DNA construct is used, the H2A C-terminal tail stabilizes the DNA ends better (in comparison with the shorter constructs), thus shifting the equilibrium towards a 'closed' nucleosome conformation even in presence of SOX transcription factor.

The 'Widom+1' DNA template had the following sequence: ATCG AGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGC ACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTTCCTTTGTG CGTTATTACTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCCGAT. The DNA-1b template with all of the TTGT motifs mutated had the following sequence: ATCTACACGACGCTCTTCCGATCTAATTT ATTCAGACTAGCGTTATACTATTCTAATTTCAGACTTCGGTGGTCAGAC TTATCAGACTCTCAGACGCGTTCAGCTTAATGCCTAACGACACTCGGAG ATCGGAAGAGCACACGTGAT.

A plasmid pMK containing four consecutive copies of a DNA template of interest separated by EcoRV restriction sites was ordered from GeneArt (Thermo Fisher). The plasmid was produced in large quantities in *E. coli* XL1 blue cells, and purified with a NucleoBond PC 10000 kit (Macherey Nagel). The plasmid was digested with EcoRV enzyme (NEB) overnight, and produced four copies of the insert per plasmid. The plasmid was then precipitated with PEG-6000[37]. The insert was further purified by size-exclusion chromatography with a Superose 6 Increase column (GE Healthcare). Peak fractions were pooled and concentrated by ethanol precipitation. Alternatively, DNA templates (DNA-1b, Widom+1 and Cy3-labelled DNA-1) were amplified from a plasmid with one insert copy via PCR. PCR product was purified via anion exchange on Resource Q 6 ml column (GE Healthcare).

Sample preparation and experiments for Förster resonance energy transfer

H2A(K119C) was prepared as described above and labelled according to ref. [38] with Cy5–maleimide (GE GEPA25031). Fluorescent DNA-1 template was produced by PCR with a Cy3 5′ labelled primer (IDT). Nucleosomes containing both labelled or just the donor Cy3 as control were reconstituted. Cy3 label was located at the SHL +7 end of the nucleosome. In the buffer containing 10 mM HEPES pH 7.5, 1 mM $MgCl_2$, 0.01 mM $ZnCl_2$, 1 mM DTT, 10 mM NaCl, 0.5 mg/ml BSA, nucleosomes (60 nM concentration) and SOX were mixed on ice and the spectra were recorded. Excitation wavelength of 510 nm was used. Förster resonance energy transfer (FRET) efficiency was calculated using a standard formula $E = 1 - I_{DA}/I_D$. Four independent experiments were performed.

Nucleosome reconstitution

Nucleosomes were reconstituted from the histone octamer and DNA template with a salt gradient as previously described[35]. In brief, octamer and DNA were mixed in 1.2:1 ratio in RB high, transferred into Slide-A-Lyzer MINI Dialysis Units 7,000 MWCO (Thermo Fisher), and dialysed gradually over a course of 24 h from RB high into RB low (20 mM HEPES pH 7.5, 1 mM EDTA, 20 mM NaCl, 2 mM DTT). Freshly reconstituted nucleosomes were concentrated in Amicon Ultra-0.5 centrifugal filters MWCO 10,000 (Sigma Aldrich).

Cryo-EM grid preparation and data collection

Nucleosomes at 1.6 μM concentration were mixed with 20× molar excess of SOX transcription factor at 4 °C in the final buffer containing 20 mM HEPES pH 7.5, 1 mM EDTA, 30 mM NaCl, 2 mM DTT, and used for cryo-grid preparation. First, R 2/1 Cu 300 mesh grids (Quantifoil) were glow-discharged with PELCO easiGlow (Ted Pella) device for 120 s. Next, 3.5 μl of sample was applied to the grid in the Vitrobot Mark IV (FEI) chamber at 100% humidity and 16 °C. The excess of liquid was blotted away for 10 s, and the grid was vitrified by plunging into liquid ethane. Data collection was performed on a G2 Titan Krios microscope (FEI) equipped with a K2 Summit direct electron detector (Gatan). Data were collected with EPU software (Thermo Fisher), with defocus ranging

from 0.9 to 3.4 µm at a nominal magnification of 130,000× and a pixel size of 1.05 Å/pixel. Data were collected with an energy filter slit set to 30 eV. The total electron dose of 45 e^-/A^2 was distributed over 40 movie frames. For all imaged samples at least 50% of the data were collected at 25° stage tilt to partially compensate for preferred orientation of particles on the grid, and to improve angular distribution. The quality of the reconstructions was improved compared to the zero-tilt data. Data collection was monitored on-the-fly with Warp[39] and cryoSPARC 2D classification[40].

Data processing and analysis

Processing details are summarized in Extended Data Table 1. For every dataset, particles were picked with gAutomatch, CTF determination was performed with Gctf[41]. The initial reference from the free-nucleosome set was obtained ab initio in cryoSPARC, low-pass-filtered to 40 Å and used as a starting point for the 3D classification of all datasets. For every dataset, to speed up the computation, binned particles with the pixel size of 4.2 Å were extracted and subjected to several rounds of 2D classification and 3D classification in Relion[42]. Classes showing high-resolution features were selected for further processing. Next, selected particles were re-extracted with a pixel size of 2.1 Å, and were 3D-classified and cleaned again. Finally, particles were re-extracted at the final pixel size of 1.05 Å and box size of 400 pixels, and subjected to 3D refinement. For all datasets, processing was performed without symmetry application (C_1). Final Fourier shell correlation (FSC) curves supplied with directional FSC curves and anisotropy estimates were calculated using 3DFSC server[43] (Extended Data Fig. 4). In addition, for each map local resolution was calculated in Relion (Extended Data Fig. 4).

For the free nucleosome dataset, after CTF refinement and 3D refinement, final maps were sharpened using *B*-factor of −75. The final dataset contained 368,270 particles. When classified, this dataset showed typical levels of partial DNA unwrapping (about 10 bp) at the nucleosome entry or exit sites in around 15% of the data (similarly to ref. [26]); however, the overwhelming majority of particles contributed to a fully wrapped nucleosome reconstruction.

In case of the nucleosome–SOX2 dataset, classes that showed additional densities were selected after 3D classification (with global soft mask applied). Next, a selected subset was subjected to a round of focused classification with a small soft spherical mask centred at the additional density near SHL +2 of the nucleosome. A class showing strong additional density was selected and further refined. The final dataset was CTF-refined to compensate for local defocus variations. As a final step, the dataset was subjected to non-uniform refinement in cryoSPARC, which led to an improved local resolution distribution in the 3D reconstruction. The final map was sharpened using a *B*-factor of −100. The final dataset contained 32,301 particles. For an overview of the processing pipelines for both nucleosome–SOX11 datasets, see Extended Data Figs. 9, 10. The starting number of particles (several million) was similar for nucleosome–SOX2 and for nucleosome–SOX11 datasets; however, nucleosome–SOX2 yielded a smaller number of 'good' particles that resulted in a final reconstruction.

The nucleosome–SOX11 dataset was processed in a similar way. The final dataset after initial steps of coarse cleaning was classified into four classes, out of which two were of high quality. One of the classes (202,142 particles) showed a clear additional density at SHL +2 and detached terminal DNA. The corresponding final map was sharpened using a *B*-factor of −100. Another class with two additional densities (nucleosome–$SOX11_2$) contained 114,104 particles. It was refined and sharpened using a *B*-factor of −120. In this nucleosome–$SOX11_2$ structure, the SOX factor molecules are located at SHL +2 and SHL −2, but are not related by the two-fold pseudo-symmetry of the nucleosome. This confirms that the density for the second SOX factor is not an artefact of particle misalignment during data processing. Lower occupancy of SOX11 at SHL −2 may be due to the presence of a weaker binding motif TTCT in that position. The local curvature induced by SOX binding at SHL −2 is not as pronounced as at SHL +2, possibly also owing to a weaker binding motif.

The nucleosome–SOX11* dataset resulted in two distinct classes. The first class (130,870 particles) resulted in a map virtually identical to the nucleosome–SOX structure, but with slightly lower resolution (about 4.0 Å). The remaining 63,821 particles resulted in the nucleosome–SOX* map. The final nucleosome–SOX* map was sharpened using *B*-factor of −100.

Model fitting and refinement

To model the free nucleosome structure, we started from a canonical nucleosome structure obtained by cryo-EM (PDB code 6FQ5)[44] and altered the DNA sequence to correspond to the DNA-1 template using Chimera[45]. Several amino acid residues in the *Xenopus laevis* histones were substituted with ones corresponding to the *H. sapiens* histones in Coot[46]. Next, the model was fitted into the corresponding sharpened cryo-EM map of the free nucleosome and refined in real space using Phenix[47].

The refined model of the free nucleosome was used to generate models for the nucleosome–SOX complex structures. In case of the nucleosome–SOX2 complex, both the nucleosome model and the SOX2 structure (PDB code 1O4X) were placed into the cryo-EM map, nucleosome DNA regions outside of the map were removed and the model was refined in real space using Phenix. For the nucleosome–SOX11 models, the nucleosome and the X-ray structure of SOX11 (determined in this study) were placed into the density and refined in real space using Phenix. In both cases, extra reference model restraints ($\sigma = 1$) were imposed to keep the model close to the available higher-resolution X-ray structure. In addition, base-pair and base-stacking restraints were used during refinements, excluding the region near the SOX transcription-factor binding site because strong local DNA distortion was evident in this region of the map. An equivalent procedure was used for modelling the other structures described here.

Electrophoretic mobility shift assay

Nucleosomes at a final concentration of 1.1 nM were mixed with purified proteins (SOX2 or SOX11 DBD). The final buffer contained 10 mM HEPES pH 7.5, 1 mM $MgCl_2$, 0.01 mM $ZnCl_2$, 1 mM DTT, 10 mM NaCl, 0.5 mg/ml BSA, 5% glycerol as in ref. [12] (Extended Data Fig. 3). Samples were incubated at 10 min at room temperature, mixed with Novex Hi-Density TBE sample buffer (Thermo Fisher), and loaded onto a 6% TBE PAGE. Electrophoresis was performed at 4 °C at 100 V in 0.5× TBE buffer for 1.5–2 h. Gels were stained with SYBR Gold dye (Thermo Fisher), washed, and imaged with Typhoon 9500 FLA Imager (GE Healthcare Life Sciences).

Electrophoretic mobility shift assays (EMSAs) (Extended Data Fig. 6) were performed identically to the procedure described above, but with higher final glycerol concentration to better observe the effects of point mutations of SOX11 on the nucleosome-binding properties of SOX11 in a wider range of apparent affinities. A control EMSA in the 12% glycerol buffer is shown in Extended Data Fig. 6a, b.

Digestion assay

Two hundred and fifty nanograms of nucleosome or DNA was mixed on ice with increasing amounts of SOX11 in digestion buffer (20 mM HEPES pH 7.5, 30 mM NaCl, 10 mM magnesium acetate, 0.1 mg/ml BSA). Then, 0.125 units of restriction enzyme BfuCI (NEB) was added to each reaction. Samples were incubated at 37 °C for 30 min, and the enzyme was inactivated by incubating at 65 °C for 20 min. Samples were then incubated with proteinase K and urea, and then were loaded onto a 4–20% TBE-gel. Electrophoresis was performed at 180 V in 1× TBE buffer for 40 min. Gels were stained with SYBR Gold dye (Thermo Fisher), washed, and imaged with a Typhoon 9500 FLA Imager (GE Healthcare Life Sciences). Two independent experiments were performed both for the DNA and nucleosome digestion assays. Band intensities for the digestion product were measured in ImageJ according to standard routine[48].

Crystallization and X-ray structure determination

DNA oligonucleotides (TATTGTTTATTTTGTT and AACAAAATAA ACAATA) were synthesized and PAGE-purified by IDT. Complimentary oligonucleotides were annealed by heating to 95 °C and stepwise cooling to 4 °C (1° per 90 s) at a concentration of 1.5 mM. Concentrated purified SOX11 DBD and 16-mer DNA were mixed in 1:1.2 ratio and incubated on ice for 30 min. Crystallization was achieved by the hanging drop vapour diffusion method at 20 °C by mixing 1 µl of sample solution with 1 µl of reservoir solution containing 100 mM NaOAc pH 4.5, 200 mM CaOAc, 17% PEG400. Crystals were cryo-protected by 35% PEG400 (v/v) in the final storage solution and flash-frozen in liquid nitrogen.

X-ray diffraction data were collected at beamline X10SA at the Swiss Light Source using a Pilatus 6M detector. Data were indexed and integrated using XDS and scaled using XSCALE[49]. The structure was solved by molecular replacement with PHASER[50], using the structure of the free SOX2 (PDB code 1GT0[51]) as the search model. The crystals belonged to space group $P6_1$ and diffracted to a resolution of 2.5 Å. The asymmetric unit contained two protein–DNA complexes (Extended Data Fig. 2d). Density modification and model building was carried out with phenix.autobuild and manually completed in Coot. The model was iteratively refined with phenix.refine and outliers were fixed in Coot. The final R_{free} factor was 26%. The final model contained SOX11 residues 46–122 and DNA nucleotides 1–14. Diffraction data and refinement statistics are summarized in Extended Data Table 2.

Estimation of the effect of Mg^{2+} on the nucleosome–SOX11 structure

Because nucleosomes are known to be sensitive to Mg^{2+} concentration, we wanted to test whether Mg^{2+} affects the nucleosome–SOX11 structure. Nucleosomes at 1.6 µM concentration were mixed with 20× molar excess of SOX11 transcription factor at 4 °C in the final buffer containing 20 mM HEPES pH 7.5, 30 mM NaCl, 1 mM DTT supplied with additional 1 mM $MgCl_2$, 0.01 mM $ZnCl_2$, 0.5% glycerol. Such sample buffer resembles the buffer used for our EMSAs except for glycerol and BSA, which should be avoided in cryo-EM samples. Next, we used this sample for cryo-grid preparation. We collected a dataset on the Titan Krios equipped with a K3 Gatan detector, nominal pixel size 1.07 Å. Processing was done similarly to the other datasets described here. The final set contained 93493 particles. After light 3D classification (removing low resolution classes), the cryo-EM map (at 4 Å resolution, with 0.73 sphericity) looked highly similar to our original nucleosome–SOX11 structure. The comparison is shown in the Extended Data Fig. 6g. We concluded that Mg^{2+} ions do not alter nucleosome–SOX structure. Overall, the Mg^{2+} sample looked better than the original one in terms of SOX occupancy and quality: a higher portion of particles from the original set contributed to the final reconstruction. We did not further analyse this dataset.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The cryo-EM density reconstructions and final models have been deposited with the Electron Microscopy Data Bank (EMD-10390, EMD-10391, EMD-10392, EMD-10393 and EMD-10394) and with the Protein Data Bank (PDB) (6T78, 6T79, 6T7A, 6T7B, 6T7C and 6T7D). All data are available in the Article and its Supplementary Information.

34. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Expression and purification of recombinant histones and nucleosome reconstitution. *Methods Mol. Biol.* **119**, 1–16 (1999).
35. Dyer, P. N. et al. Reconstitution of nucleosome core particles from recombinant histones and DNA. *Methods Enzymol.* **375**, 23–44 (2004).
36. Li, Z. & Kono, H. Distinct roles of histone H3 and H2A tails in nucleosome stability. *Sci. Rep.* **6**, 31437 (2016).
37. Lis, J. T. & Schleif, R. Size fractionation of double-stranded DNA by precipitation with polyethylene glycol. *Nucleic Acids Res.* **2**, 383–390 (1975).
38. Shimko, J. C., North, J. A., Bruns, A. N., Poirier, M. G. & Ottesen, J. J. Preparation of fully synthetic histone H3 reveals that acetyl-lysine 56 facilitates protein binding within nucleosomes. *J. Mol. Biol.* **408**, 187–204 (2011).
39. Tegunov, D. & Cramer, P. Real-time cryo-electron microscopy data preprocessing with Warp. *Nat. Methods* **16**, 1146–1152 (2019).
40. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
41. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
42. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
43. Tan, Y. Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).
44. Bilokapic, S., Strauss, M. & Halic, M. Structural rearrangements of the histone octamer translocate DNA. *Nat. Commun.* **9**, 1330 (2018).
45. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
46. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
47. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
48. Rueden, C. T. et al. ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529 (2017).
49. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
50. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D* **63**, 32–41 (2007).
51. Reményi, A. et al. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* **17**, 2048–2059 (2003).
52. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
53. Klaus, M. et al. Structure and decoy-mediated inhibition of the SOX18/Prox1-DNA interaction. *Nucleic Acids Res.* **44**, 3922–3935 (2016).
54. Palasingam, P., Jauch, R., Ng, C. K. & Kolatkar, P. R. The structure of Sox17 bound to DNA reveals a conserved bending topology but selective protein interaction platforms. *J. Mol. Biol.* **388**, 619–630 (2009).
55. Werner, M. H., Huth, J. R., Gronenborn, A. M. & Clore, G. M. Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY–DNA complex. *Cell* **81**, 705–714 (1995).

Acknowledgements We thank members of the Cramer laboratory, in particular H. Hillen, D. Tegunov, and G. Kokic for advice; the crystallization facility at our institute, in particular J. Wawrzinek and U. Steuerwald; and W. Fischle for providing histone expression constructs. Part of this work was performed at Beamline X10SA at the SLS at the PSI. S.O.D. was supported by an EMBO long-term fellowship (ALTF-949-2016). P.C. was supported by the Deutsche Forschungsgemeinschaft (EXC 2067/1-390729940), the ERC Advanced Investigator Grant TRANSREGULON (grant agreement no 693023) and the Volkswagen Foundation.

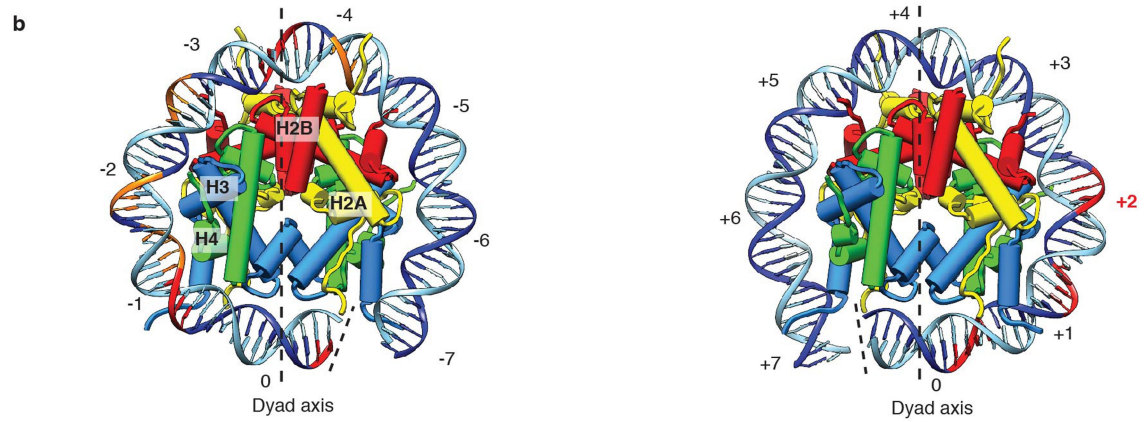
Author contributions S.O.D. designed and carried out all experiments and data analysis. F.Z., supported by J.T., identified the original DNA template used in the study. C.D. assisted with cryo-EM data collection. P.C. designed and supervised research. S.O.D. and P.C. interpreted the data and wrote the manuscript, with input from all authors.

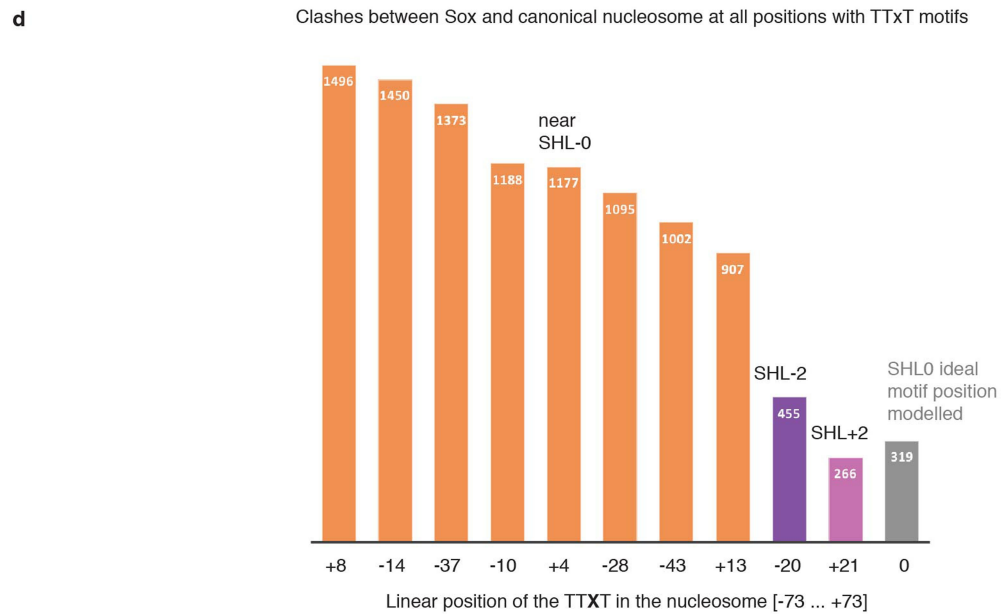
Competing interests The authors declare no competing interests.

Additional information

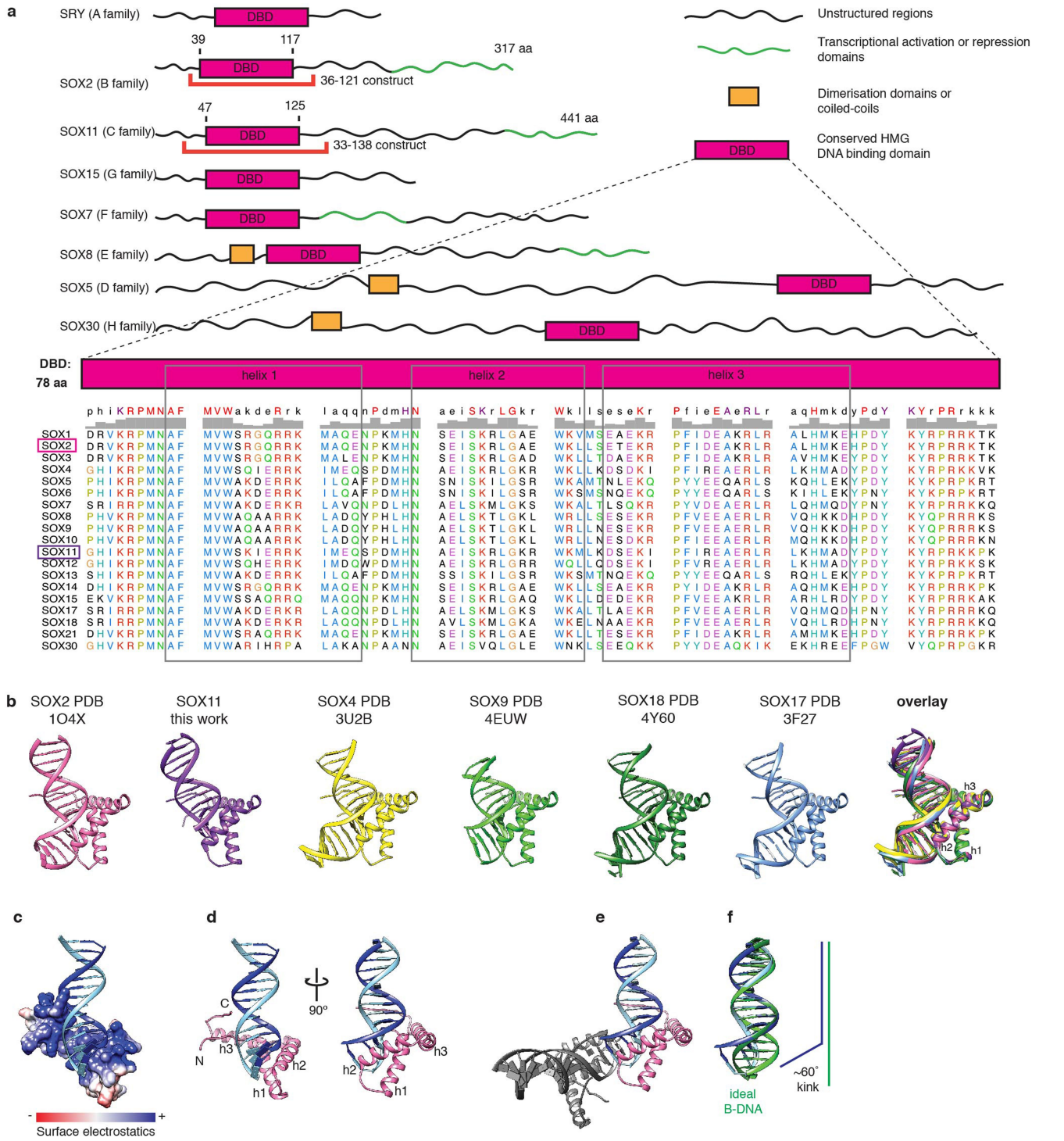
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2195-y.

Correspondence and requests for materials should be addressed to P.C.

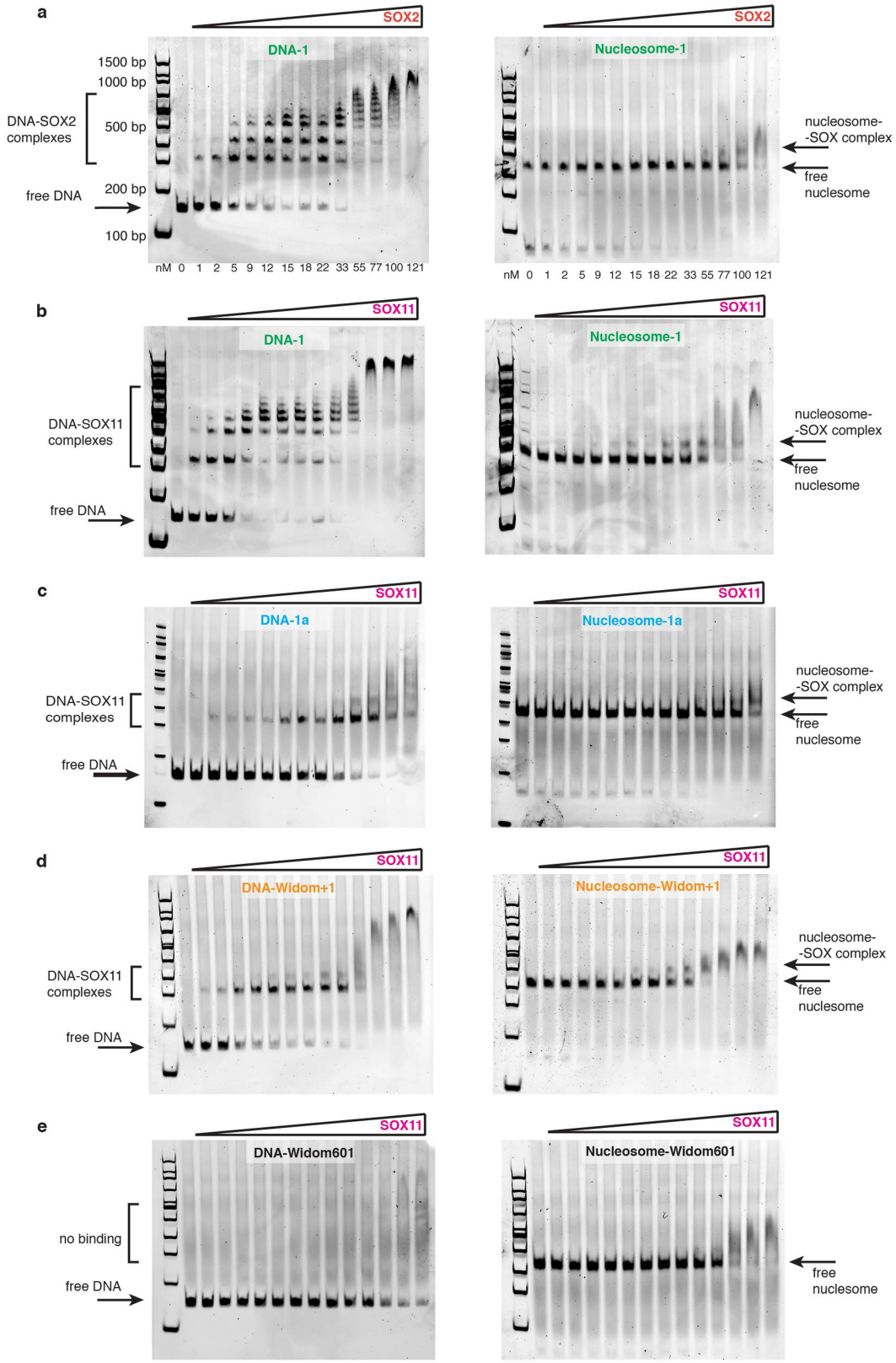




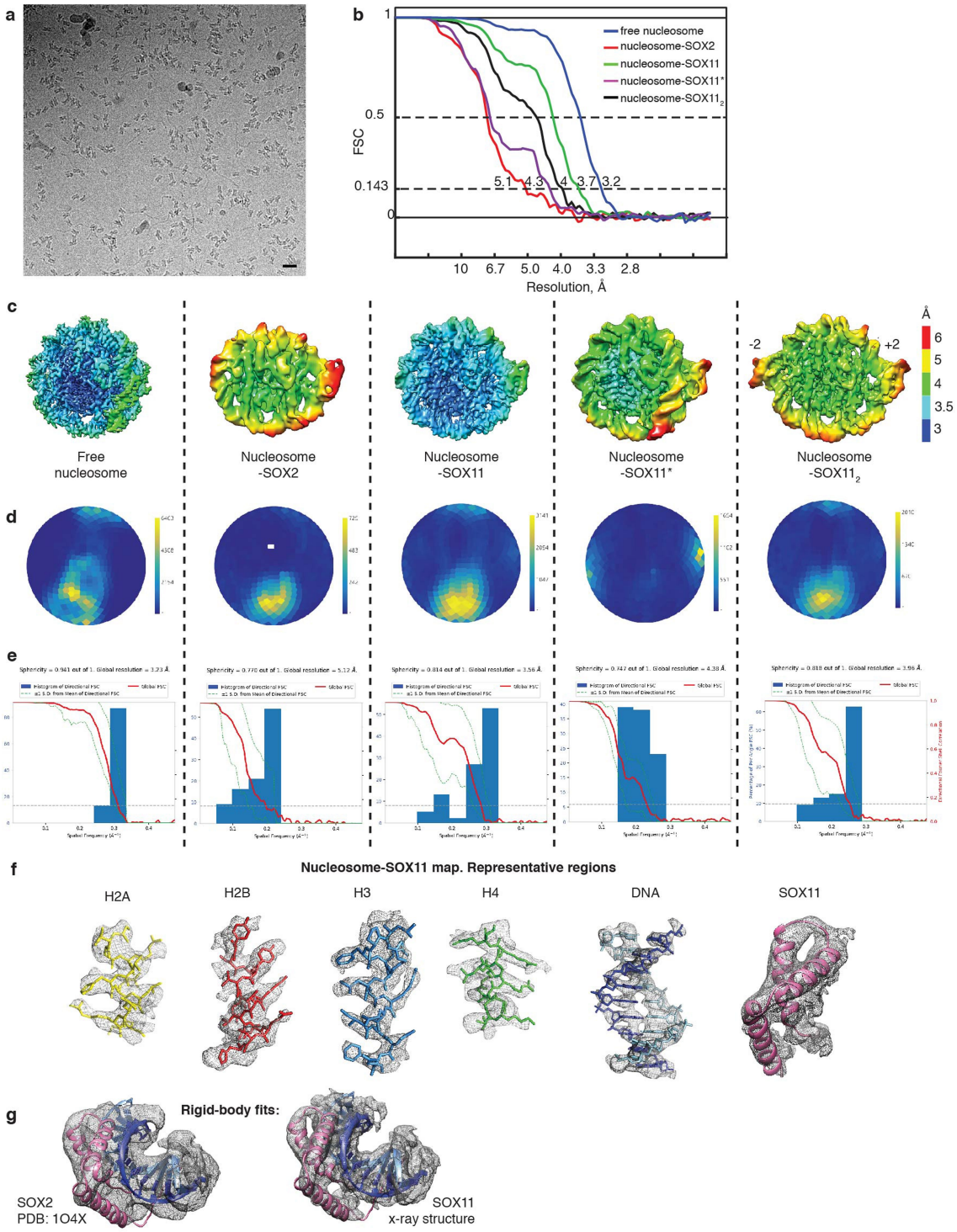
Extended Data Fig. 1 | DNA constructs and motif positions. Related to Fig. 1. **a**, DNA-1 sequence. Two DNA strands are coloured in dark and light blue, canonical core motifs TTGT are coloured in red, TTXT motifs are shown in orange. Only motifs that allow SOX binding to the DNA minor groove are considered. The position of the third nucleotide of each motif in the DNA-1 sequence is indicated. Motifs at SHL+2 and −2 are shown in bold. **b**, Top views of the nucleosome. H2A, H2B, H3 and H4 are shown in yellow, red, blue and green, respectively, DNA is shown in dark and light blue. SHLs are labelled. **c**, DNA-1a template sequence. Only one TTGT motif is present (red). **d**, Structure of SOX2 (PDB code 1O4X) was aligned to each of the motifs present in nucleosome 1 and allowing binding of SOX to the minor groove. The number of clashes was calculated using 'findclash' command in Chimera software. SOX2 binding to motifs at SHL +2 and SHL −2 gives rise to the least amount of clashes with DNA and histones compared to other locations. The ideal position (modelled) of the SOX motif on the dyad would result in a comparably low number of clashes.



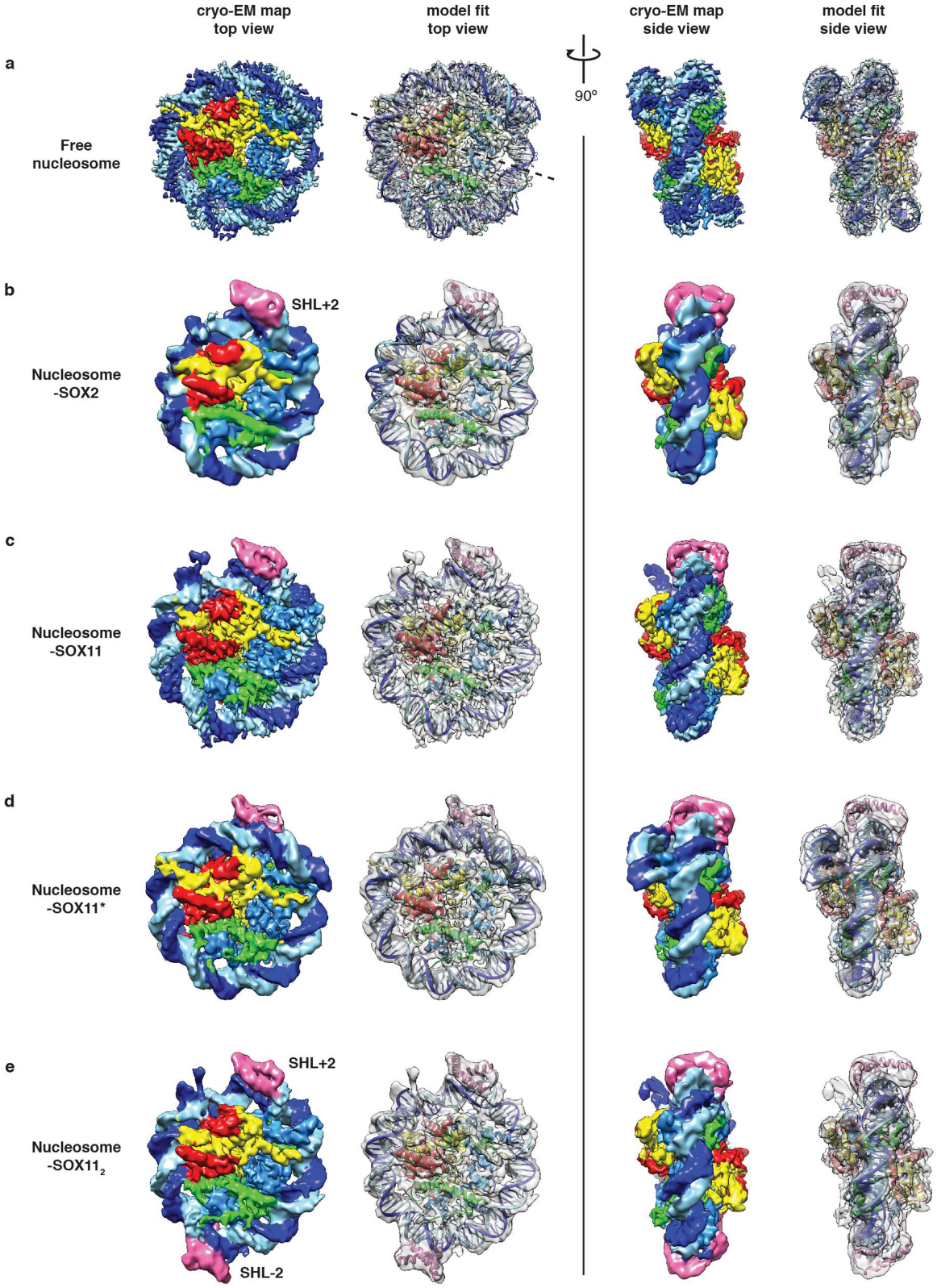
Extended Data Fig. 2 | Conservation of SOX-family DBD sequence and structure, and X-ray structure of the SOX11–DNA complex. Related to Fig. 1. **a**, Domain organization of the human SOX protein family. DBDs are shown as pink rectangles; unstructured functionally diverse regions are shown as wavy lines. Protein constructs used in this study are marked. The alignment of DBD sequences (produced using Clustal Omega) is shown below[52]. **b**, Structural conservation of SOX factors. Crystal and nuclear magnetic resonance structures of SOX transcription factors: SOX2 (ref. [22]), SOX11 (this study), SOX4 (ref. [18]), SOX9, SOX18 (ref. [53]) and SOX17 (ref. [54]); SRY (ref. [55]) has a similar fold. Superimposition of all the structures reveals that they are virtually identical. **c**, DNA is engulfed by the strongly positively charged inner surface of the SOX11 DBD. **d**, Ribbon representation of the SOX11 X-ray structure. **e**, Two copies of SOX11–DNA in the asymmetric unit. The contact between the two is mediated by DNA stacking. **f**, Comparison of the observed DNA conformation with canonical B-DNA (green). SOX11 introduces a kink into DNA that is typical for HMG box proteins.



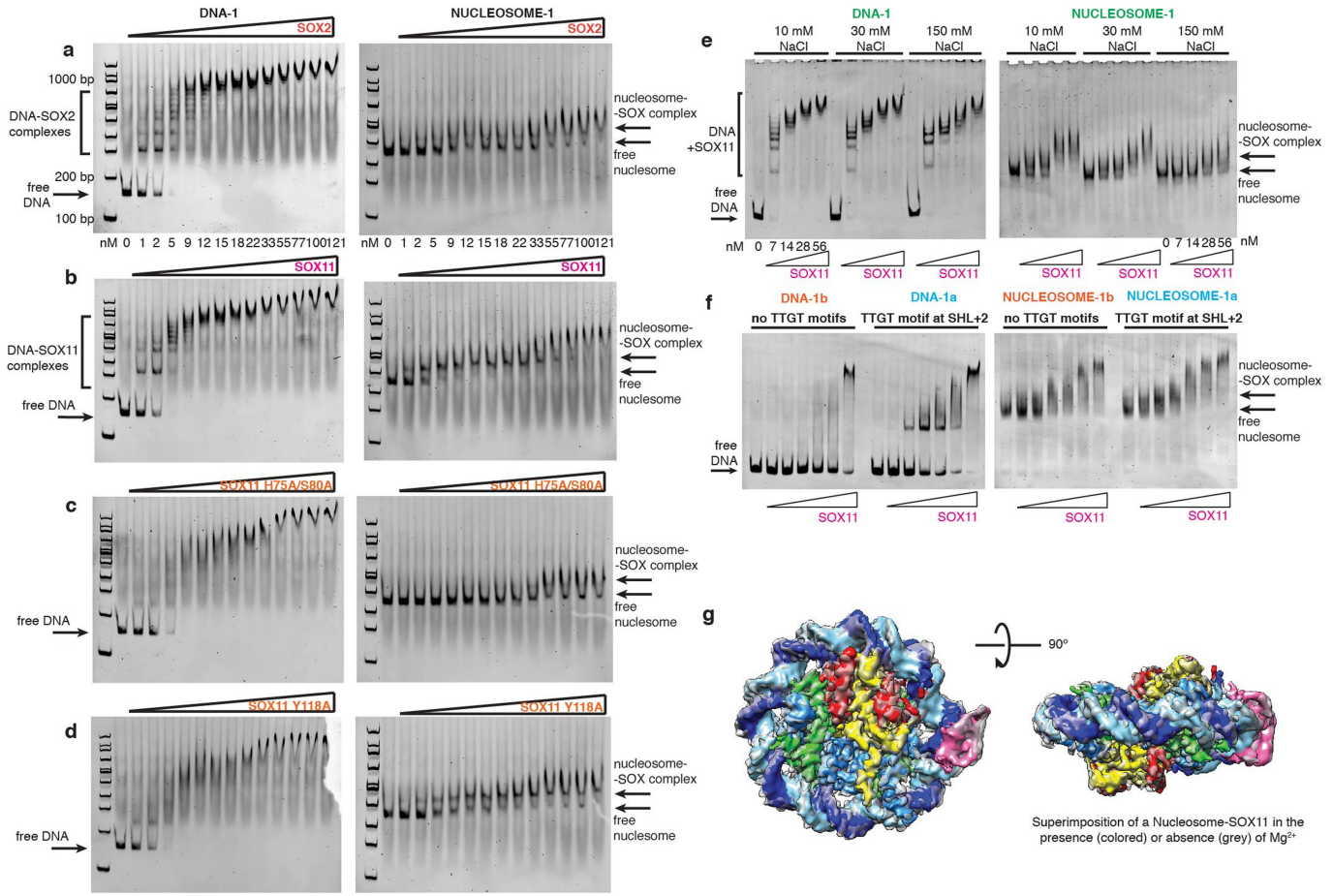
Extended Data Fig. 3 | EMSAs of SOX2 and SOX11 in complex with DNA or nucleosomes. Related to Figs. 1–3. EMSAs reveal formation of SOX-factor complexes with DNA (left) or nucleosomes (right). DNA or nucleosome concentration is 1.1 nM. **a**, EMSA of DNA-1–SOX2 and nucleosome-1–SOX2 complexes. **b**, EMSA of DNA-1–SOX11 and nucleosome-1–SOX11 complexes. **c**, EMSA of DNA-1a–SOX11 and nucleosome-1a–SOX11 complexes. **d**, EMSA of DNA–Widom+1 and nucleosome–Widom+1–SOX11 complexes. **e**, EMSA of DNA Widom 601–SOX11 and nucleosome Widom 601–SOX11 complexes. Relevant bands are labelled. All experiments were repeated at least twice with similar results. For gel source data, see Supplementary Fig. 1.



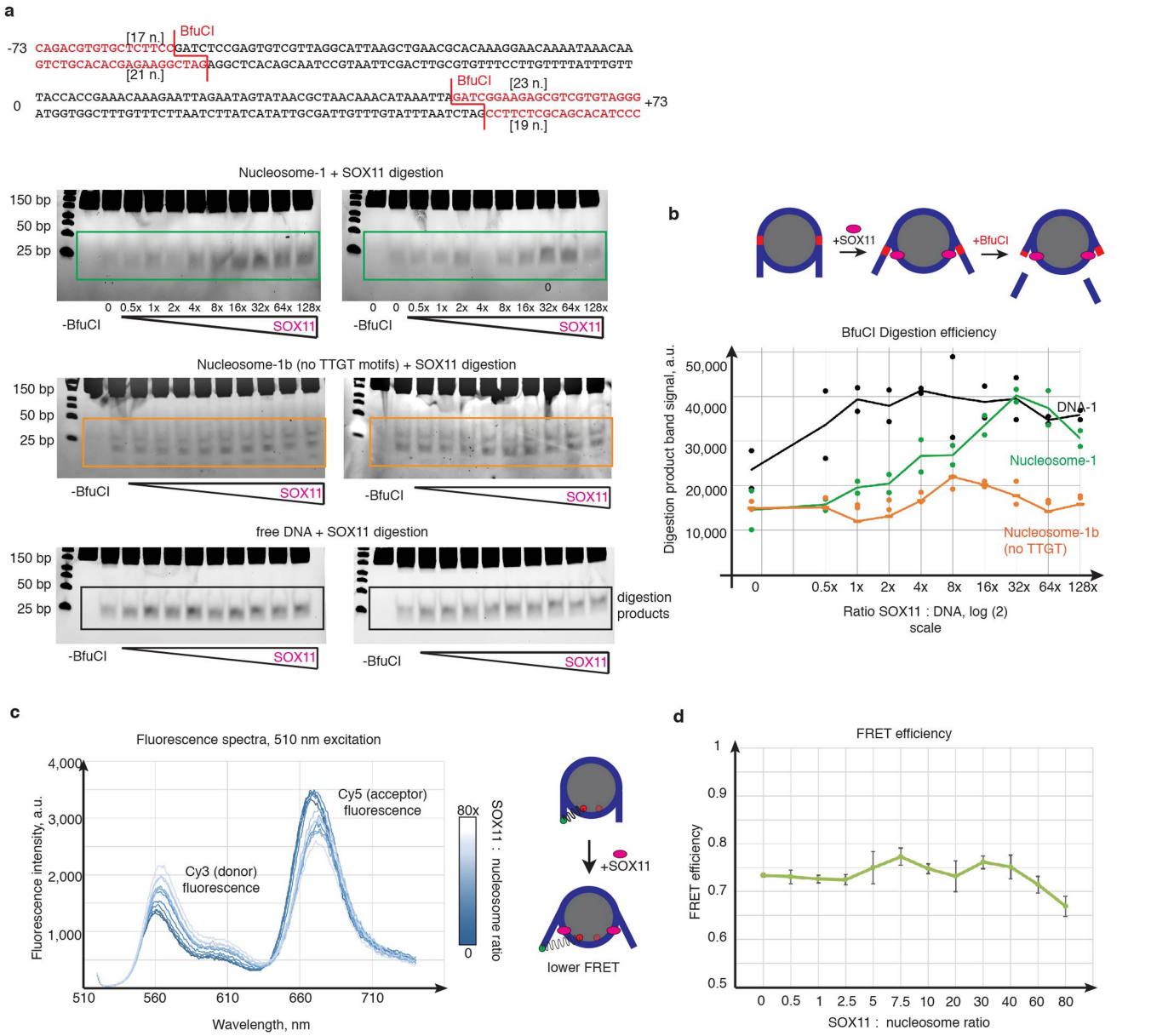
Extended Data Fig. 4 | Global and local resolution of cryo-EM reconstructions. Related to Figs. 1–3. **a**, Example micrograph reveals preferred orientation of nucleosomes. Scale bar, 20 nm. **b**, FSC plots for five reported reconstructions. **c**, Local resolution distribution. In some maps, the resolution varies from 3 Å (dark blue) through 4 Å (green) to 6 Å (red). The lower resolution of some DNA regions indicates flexibility. **d**, Angular distribution plots. Scale shows the number of particles assigned to a particular angle. Blue, a low number of particles; yellow, a high number of particles. **e**, Directional FSC plots for the reconstructions calculated on the 3DFSC server[48]. Sphericity, as the degree of anisotropy present in the reconstructions, is indicated above each plot. Histograms indicate the portion of voxels with a particular resolution. **f**, Visualization of different regions of the nucleosome–SOX11 map. In the nucleosome core, histone side chains are clearly visible; SOX density has a lower resolution, but helical densities are clearly distinguishable. **g**, Rigid-body fit of the SOX2–DNA structure (PDB code 1O4X) into the nucleosome–SOX2 cryo-EM map (left). Rigid-body fit of the SOX11–DNA X-ray structure into the nucleosome–SOX11 cryo-EM map (right). The region containing SOX and a short DNA stretch was isolated from the rest of the map for clarity.



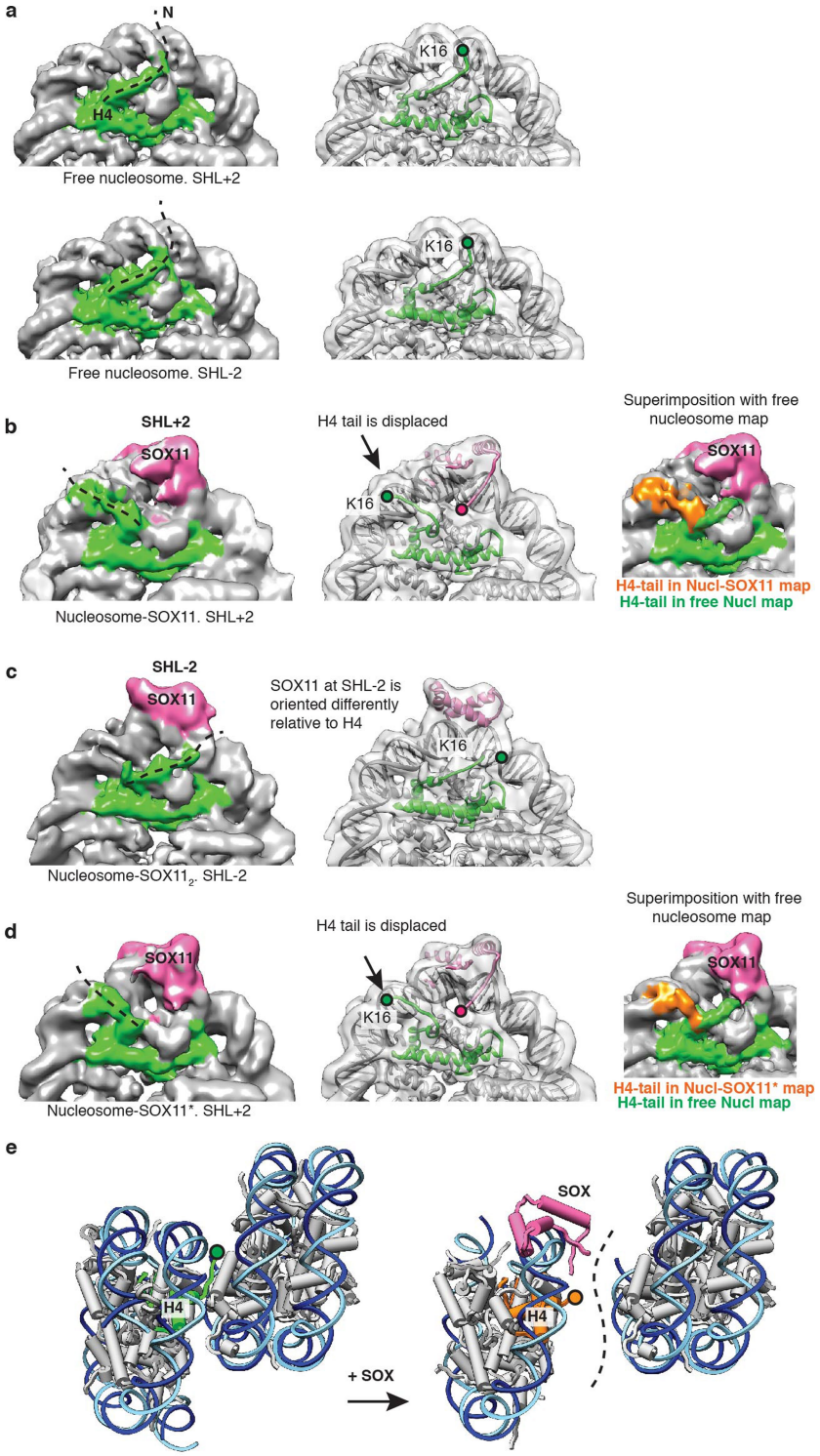
Extended Data Fig. 5 | Gallery of cryo-EM structures. Related to Figs. 1–4. **a**–**e**, Electron microscopy maps and corresponding models of all reported structures. Top views (left) and side views (right) are related by a 90° rotation. The maps are coloured on the basis of the fitted model (as in Fig. 1), or are transparent.



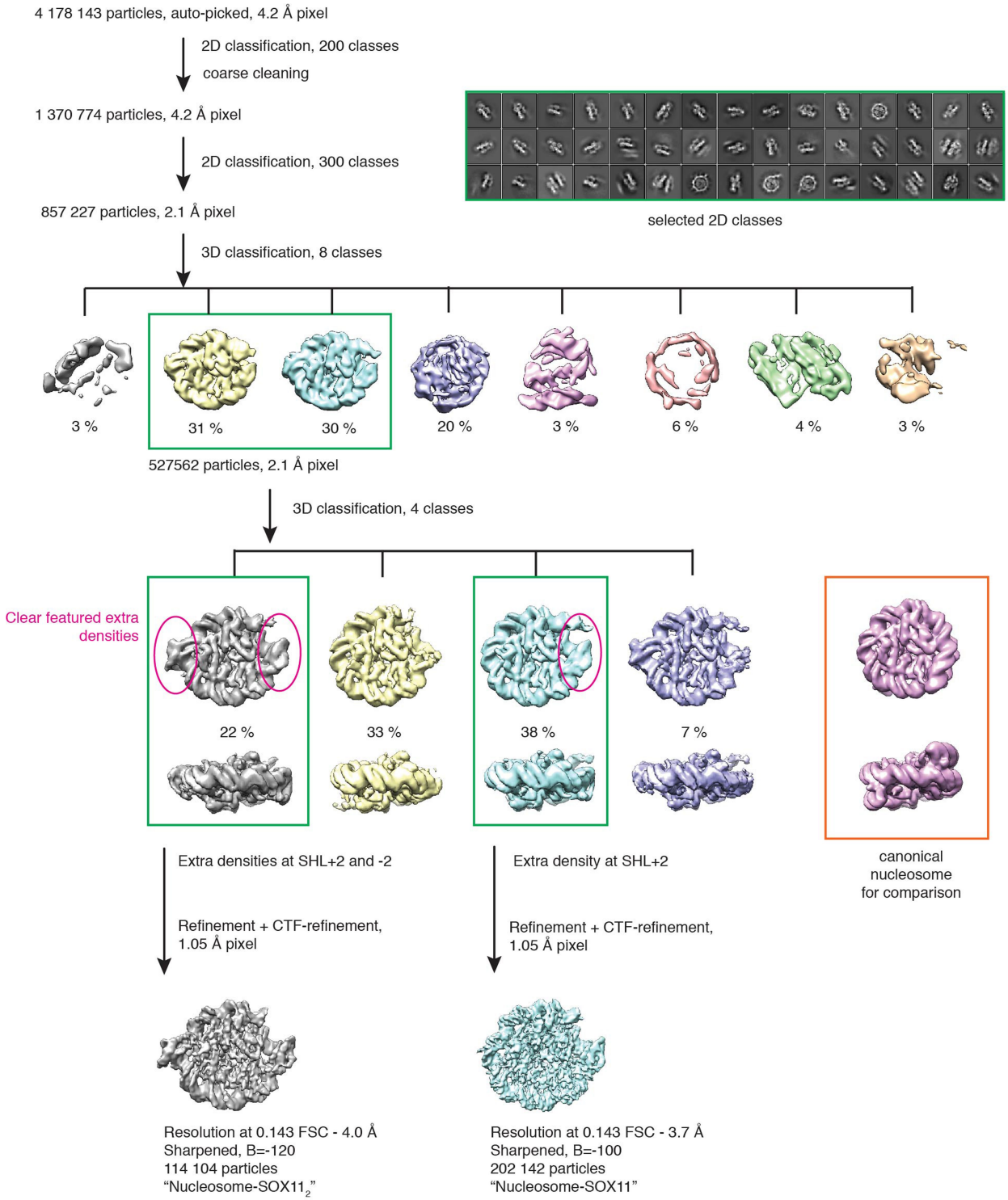
Extended Data Fig. 6 | EMSAs using SOX2, and wild-type or mutant SOX11 in complex with DNA or nucleosomes. Related to Figs. 1–3. EMSAs reveal the formation of SOX-factor complexes with DNA or nucleosomes. All experiments were repeated at least twice with similar results. For gel source data, see Supplementary Fig. 1. **a**, EMSA of DNA-1 or nucleosome-1 with wild-type SOX2. **b**, EMSA of DNA-1 or nucleosome-1 with wild-type SOX11. **c**, EMSA of DNA-1 or nucleosome-1 with SOX11(H75A/S80A). **d**, EMSA of DNA-1 or nucleosome-1 with SOX11(Y118A). Relevant bands are labelled. To observe a wider range of the binding curve for the mutants, a larger amount of glycerol (12% final concentration) was used here (as compared to the EMSAs shown in Extended Data Fig. 3, which used 5% glycerol)—thus, the apparent affinity is higher. DNA or nucleosome concentration is 1.1 nM. **e**, EMSAs reveal formation of SOX–DNA or SOX–nucleosome complexes at different concentrations of salt. There is virtually no difference in binding at 10 or 30 mM NaCl, whereas binding is weaker at 150 mM NaCl. DNA or nucleosome concentration is 1.1 nM. **f**, EMSAs reveal formation of SOX complexes with DNA-1a or nucleosome-1a (only one canonical motif present) as compared to DNA-1b and nucleosome-1b (in which all canonical SOX motifs were mutated). SOX concentrations ranged from 0 to 200 nM, DNA concentration was 1.1 nM. **g**, Superimposition of two nucleosome–SOX11 cryo-EM maps obtained in the presence of 1 mM $MgCl_2$ (grey density) or in the absence of $MgCl_2$ (coloured). Magnesium does not influence the structure of the SOX–nucleosome complex. Cross-correlation between the two unmasked maps is 0.94 (Chimera).



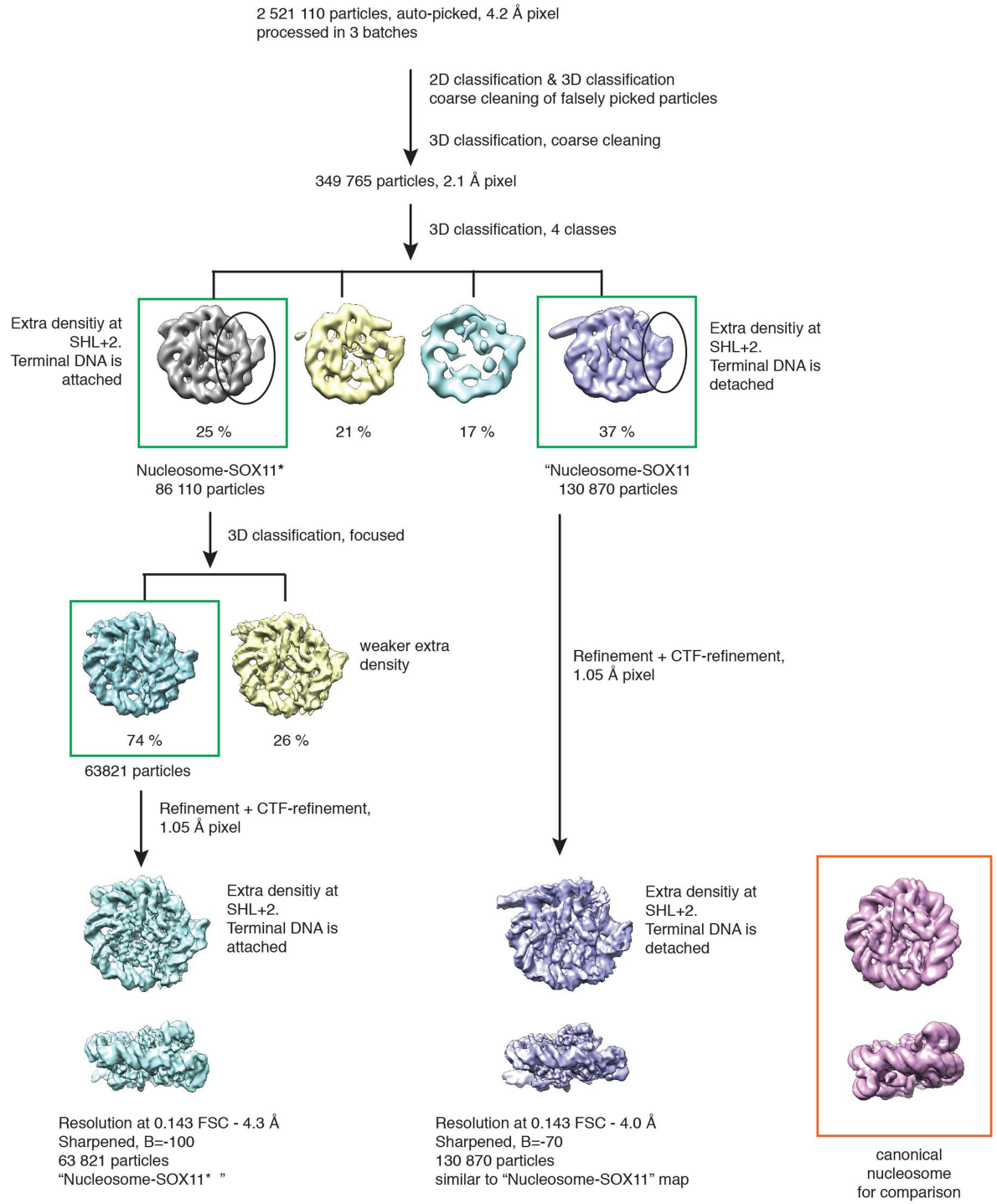
Extended Data Fig. 7 | Nucleosome-DNA end unwrapping. Related to Figs. 1, 3. **a**, DNA-1 sequence with BfuCI digestion sites. BfuCI digestion assays (two independent experiments for each sample, shown here) of the DNA-1, nucleosome-1 or nucleosome-1b (no TTGT motifs) in the presence of increasing amounts of SOX11. The restriction site (about 20 nucleotides away from the entry and exit sites of the nucleosome) becomes more accessible with higher concentrations of SOX11. In a DNA-1 digestion assay, the amount of digestion product increases slightly in the low SOX11 concentration range, and then stays constant over a broad concentration range. For gel source data, see Supplementary Fig. 1. **b**, BfuCI digestion assay plot for free DNA-1 (black), nucleosome-1 (green) and mutated nucleosome-1b (orange) in the presence of increasing amounts of SOX11. Each experiment was performed independently twice ($n = 2$). Mean values (lines) and individual measurements (dots) are shown. Band intensity was calculated using standard routine in ImageJ[52]. **c**, Example fluorescence spectra of Cy3–Cy5-labelled nucleosome in the presence of increasing amounts of SOX11. **d**, FRET efficiency plot. Mean values with s.d. are shown (independent experiments, $n = 4$).



Extended Data Fig. 8 | Repositioning of the H4 N-terminal tail. Related to Fig. 4. Cryo-EM maps are shown with a Gaussian smoothening filter (Chimera[49]) applied for clarity. SOX is coloured in pink, H4 is shown in green and the repositioned H4 tail is shown in orange. **a**, Free nucleosome. Views of SHL +2 and SHL −2 are shown to illustrate the position of the H4 tail (a dashed line). Residue K16 is marked with a circle. **b**, Nucleosome–SOX11 with SOX11 located at SHL +2. On the right, a superimposition with the free nucleosome map is shown to highlight different orientations of the H4 tail. **c**, Nucleosome–SOX11 complex with SOX11 located at SHL −2. In this location, SOX is oriented differently and does not clash with or reposition the H4 tail. **d**, Nucleosome–SOX11*. The H4 N-terminal tail is repositioned compared with that in the free nucleosome. Repositioning of the H4 tail was reported in case of strong distortions in the nucleosome structure[30]. **e**, SOX binding repositions the H4 N-terminal tail and might impair nucleosome stacking. Side view of two stacking nucleosomes (from PDB code 1AOI). H4 interacts with the acidic patch on the neighbouring H2A–H2B histone dimer. H4 tail repositioning is incompatible with nucleosome stacking.



Extended Data Fig. 9 | Flow chart for the determination of the cryo-EM structure of the nucleosome–SOX11$_2$ and nucleosome–SOX11 complexes with 147-bp DNA-1. Related to Figs. 1–3. The processing chart for the 147-bp DNA-1 nucleosome and SOX11 is depicted. The two resulting structures are nucleosome–SOX11$_2$ and nucleosome–SOX11.



Extended Data Fig. 10 | Flow chart for determination of the cryo-EM structure of the nucleosome–SOX11* complex with 151-bp DNA-1. Related to Figs. 1–3. The processing chart for the 151-bp DNA-1 nucleosome and SOX11 is depicted. The two resulting structures are nucleosome–SOX11* and a map virtually identical to the nucleosome–SOX11 from the 147-bp sample.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	Free nucleosome (EMD-10390) (PDB 6T79)	Nucleosome-SOX2 (EMD-10392) (PDB 6T7B)	Nucleosome-SOX11* (EMD-10394) (PDB 6T7D)	Nucleosome-SOX11 (EMD-10391) (PDB 6T7A)	Nucleosome-$SOX11_2$ (EMD-10393) (PDB 6T7C)
Data collection and processing					
Magnification	130 000x	130 000x	130 000x	130 000x	130 000x
Voltage (kV)	300	300	300	300	300
Electron exposure (e–/Å^2)	45	45	45	45	45
Defocus range (μm)	0.9-3.4	0.9-3.4	0.9-3.4	0.9-3.4	0.9-3.4
Pixel size (Å)	1.05	1.05	1.05	1.05	1.05
Symmetry imposed	C1	C1	C1	C1	C1
Initial particle images (no.)	451 990	2 347 743	2 521 110	4 178 143	4 178 143
Final particle images (no.)	368 270	32 301	63 821	202 142	114 104
Map resolution (Å) 0.143 FSC threshold	3.2	5.1	4.3	3.7	4.0
Map resolution range (Å)	3-3.9	4.5-8.3	4.1-6.6	3.5-5.5	3.7-5.3
Refinement					
Initial model used (PDB code)	6FQ5	6T7A	6T7A	6T79	6T7A
Model resolution (Å) FSC threshold	3.2	5.1	4.3	3.7	4.0
Model resolution range (Å)	3-3.9	4.5-8.3	4.1-6.6	3.5-5.5	3.7-5.3
Map sharpening *B* factor (Å^2)	-75	-100	-100	-100	-120
Model composition					
Non-hydrogen atoms	12121	7248	8209	10338	7558
Protein residues	776	838	853	847	916
DNA	290	190	234	190	190
B factors (Å^2)					
Protein	33	76	140	89	106
DNA	81	139	210	140	140
R.m.s. deviations					
Bond lengths (Å)	0.52	0.60	055	0.52	0.56
Bond angles (°)	0.83	0.88	0.93	0.79	0.86
Validation					
MolProbity score	1.28	1.84	2.85	1.37	0.94
Clashscore	5	1	3	7	2
Poor rotamers (%)	0	na*	na*	0	na*
Ramachandran plot					
Favored (%)	99	na*	na*	98.1	na*
Allowed (%)	1	na*	na*	1.8	na*
Disallowed (%)	0	na*	na*	0.1	na*

na*, model is deposited as backbone only (owing to insufficient resolution of the cryo-EM map).

Extended Data Table 2 | Data collection and refinement statistics (molecular replacement)

	SOX11-DNA PDB 6T78
Data collection	
Space group	P 6_1
Cell dimensions	
a, *b*, *c* (Å)	106.1, 106.1, 76.9
α, β, γ (°)	90 90 120
Wavelength (Å)	1
Resolution (Å)	46 - 2.5 (2.6-2.5)
R_{merge}	0.13 (2.8)
$I / \sigma I$	17.3 (1.28)
Completeness (%)	99.7 (98.7)
Redundancy	20.5 (20.3)
Refinement	
Resolution (Å)	2.5
No. reflections	17033 (1682)
R_{work} / R_{free}	0.23 / 0.26
No. atoms	2332
Protein	1307
DNA	1019
Water	6
B-factors	96
Protein	83
DNA	114
Water	65
R.m.s. deviations	
Bond lengths (Å)	0.01
Bond angles (°)	0.8

Number of crystals = 1. Values in parentheses are for the highest-resolution shell.

nature research

Corresponding author(s): Patrick Cramer

Last updated by author(s): Mar 6, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist .

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	FEI EPU 1.10;
Data analysis	Warp 1.0.6; RELION 3.0; Gautomatch v0.56, Gctf v1.06, CryoSPARC v2; COOT v. 0.8.9.1; PHENIX v. 1.14; Pymol v. 2.2.3; Chimera v. 1.13 ; Molprobity 4.5 server; 3DFSC server; ImageJ 1.52k; Microsoft Excel 14.7.7; XSCALE BUILT=20170601; PHASER 1.2;

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The electron density reconstructions and corresponding models were deposited with the Electron Microscopy Data Base (EMD-10390, EMD-10391, EMD-10392, EMD-10393, EMD-10394) and with the Protein Data Bank (6T78, 6T79, 6T7A, 6T7B, 6T7C, 6T7D). Source data is relevant for Fig. 2, ED Figures 2,4,5,6,10,13

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Numbers of particles used for cryo-EM reconstructions are listed in Extended Table 1. Sample sizes were estimated on the basis of previous studies using similar methods and analyses that are widely published.
Data exclusions	No data were excluded from the analysis.
Replication	All attempts at replication were successful and reproducible. At least two independent biological repeats per experiment performed and showed similar results. Structure determination does not require replication.
Randomization	Samples were not allocated into groups. Randomization is not relevant to this study.
Blinding	Blinding was not relevant to this study because there was no group allocation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Matters arising

On the signature of a 70-solar-mass black hole in LB-1

https://doi.org/10.1038/s41586-020-2216-x

Received: 6 December 2019

Accepted: 27 February 2020

Published online: 29 April 2020

Michael Abdul-Masih[1], Gareth Banyard[1], Julia Bodensteiner[1], Emma Bordier[1], Dominic M. Bowman[1], Karan Dsilva[1], Matthias Fabry[1], Calum Hawcroft[1], Laurent Mahy[1], Pablo Marchant[1], Gert Raskin[1], Maddalena Reggiani[1], Tomer Shenar[1], Andrew Tkachenko[1], Hans Van Winckel[1], Lore Vermeylen[2] & Hugues Sana[1]✉

ARISING FROM: J. Liu et al. *Nature* https://doi.org/10.1038/s41586-019-1766-2 (2019)

Massive stellar-mass black holes are not expected in a Galactic metallicity environment, owing to strong stellar winds and pair-instability supernovae. In this context, the recent report[1] of an approximately 70-solar-mass ($M_\odot$) black hole in the galactic binary system LB-1 challenges conventional theories of massive-star evolution, stellar winds and core-collapse supernovae, thus requiring a more exotic scenario to explain the existence and properties of this system[2,3]. Here we show that the apparent shifts of the Hα emission line used to derive the mass of the black hole arise from the orbital motion of the B-type companion star in the LB-1 binary system and not from that of the black hole. No evidence for a massive black hole remains in the data, thus removing the existing tension between its proposed existence and models of the formation of such a massive black hole at galactic metallicity.

LB-1 is a recently discovered galactic spectroscopic binary system with a 78.9-day period. Using multi-epoch optical spectroscopy from the LAMOST and Keck telescopes, Liu et al.[1] recently reported the detection of a $68^{+11}_{-13} M_\odot$ black hole paired with an $8.2^{+0.9}_{-1.2} M_\odot$ B-type subgiant star. This black hole is over twice as massive as any other known stellar-mass black hole with non-compact companions[4,5] and its mass approaches the masses that result from mergers of black holes detected by gravitational wave interferometers[6].

The detection of a large black-hole mass relies on two main lines of evidence: (1) the characterization of the orbital and physical properties of its companion B-type-star; and (2) an indirect measurement of the reflex orbital motion of the black hole.

The critical measure that yields the high mass for the hidden companion comes from the semi-amplitude $K_{BH} = 6.4\ \mathrm{km\,s^{-1}}$ of the radial-velocity curve of this potential black hole, a result that is measured from the apparent wobbling of the position of the wings of the Hα emission profile, assuming that this emission is generated by a passive accretion disk around the unseen companion. Although Liu et al.[1] assumed that such wobbling was tracing the reflex motion of the black hole, we show here that the observed radial-velocity measurements result from the superposition of the stellar absorption line of the B-type star on a static Hα emission profile. This can be demonstrated either observationally or solely by simulation.

To reach this conclusion, we used new high-spectral-resolution observations of LB-1 obtained with the HERMES spectrograph[7] coupled with the Mercator telescope on the island of La Palma, Spain (see Supplementary Information). Using these data, we isolated the pure Hα emission profile by subtracting from the observed profile a theoretical Hα absorption profile corresponding to the best-fit atmospheric parameters of the detected B-type star (see Supplementary Information), after accounting for the orbital shift of the observed profile given the epoch of the new HERMES observations. The pure emission profile is displayed in Fig. 1. We then create a series of reconstructed Hα profiles that combine the (static) emission profile and the phase-shifted Hα absorption from the B-type-star companion at phases corresponding to those observed by Liu et al.[1] (Fig. 2). Finally, we used a barycentre method as well as a bisector method using an identical mask as that in Liu et al.[1], to estimate the apparent radial-velocity shift resulting from the combined Hα profile, obtaining similar results using both methods. The radial-velocity measurements are displayed in Fig. 2 and reveal a clear low-amplitude sinusoidal variation in perfect anti-phase with that of the B-type star. In addition, the derived (apparent) semi-amplitude of $5.4\ \mathrm{km\,s^{-1}}$ is in good agreement with that measured by Liu et al.[1] ($6.4 \pm 0.8\ \mathrm{km\,s^{-1}}$), given the higher spectral-resolving power of our spectra. Very similar results were obtained when repeating

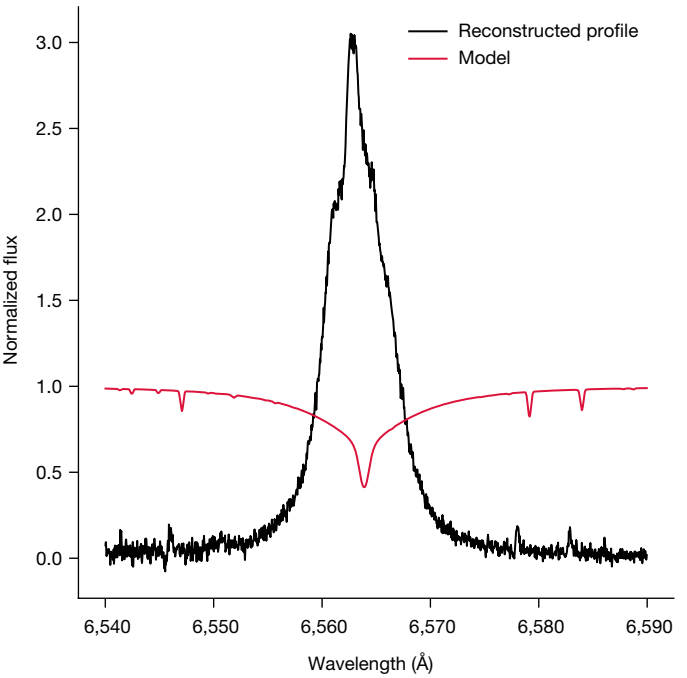


Fig. 1 | Reconstructed Hα emission profile in LB-1. Reconstructed Hα emission profile (black) obtained by subtracting the B-type star stellar absorption component (red) corresponding to the best-fit atmospheric model (see Supplementary Information) to the observed HERMES spectrum.

[1]Institute of Astronomy, KU Leuven, Leuven, Belgium. [2]Royal Observatory of Belgium, Brussels, Belgium. ✉e-mail: hugues.sana@kuleuven.be

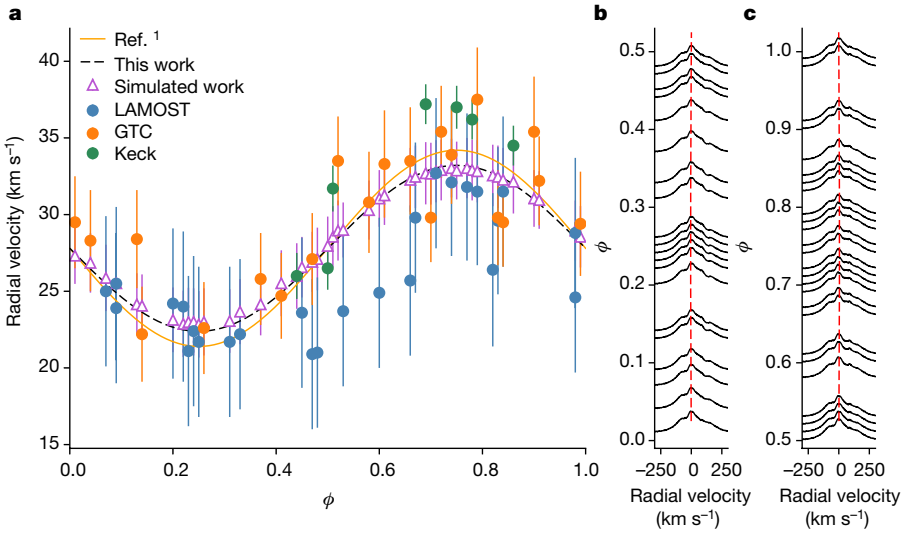


Fig. 2 | Simulating the radial-velocity signal resulting from a static Hα emission and a Doppler-shifting stellar absorption. a, Filled circles are the radial-velocity measurements and ephemeris from Liu et al.[1] and obtained by applying the barycentre method to the observed emission Hα profile. Open triangles are the results from our simulated profiles, which assume a fixed emission profile and a superimposed Hα absorption line from the B-type-star component. The best-fit radial-velocity curves from the two datasets have semi-amplitudes of 6.4 km s^{-1} (solid yellow line, orbital solution from ref. [1]) and 5.4 km s^{-1} (dashed black line, resulting from a fit to the open symbols). **b**, **c**, Simulated Hα profiles formed by a static emission component and a Doppler-shifted stellar absorption according to the orbital phase ϕ (**b**, $\phi = 0.0–0.5$; **c**, $\phi = 0.5–1.0$) and ephemeris from Liu et al.[1].

the procedure with a Gaussian profile representing the Hα line instead of the line constructed from observations (see Supplementary Information and the report[8] of a similar experiment conducted independently).

This demonstrates that the largest contribution of the detected radial-velocity signal obtained from the barycentre method applied to the Hα profile can be readily explained by a contamination of the strong Hα emission component with the absorption component of the B-type-star, and thus does not trace the reflex orbital motion of the black hole. Consequently, there is no evidence for a large mass ratio, and hence also no evidence for a large absolute mass of a black hole. The remaining observational constraints describe the minimum mass of the companion to the B-type star—of the order of $4M_\odot$ after reassessing the physical parameters of the B-type star companion (see below)—and a static or very low amplitude (<1 km s^{-1}) emission in Hα.

We also reassessed the physical parameters of the B-type-star companion by fitting local thermodynamic equilibrium (LTE) atmospheric models[8,9] to the HERMES optical spectrum (see Extended Data Figs. 1, 2) and obtained an effective temperature of $T_{\rm eff} = 13{,}500 \pm 700$ K, a surface gravity of $\log g = 3.3 \pm 0.3$ (calculated using in cgs units) and a projected rotational velocity at the equator of $v_{\rm eq}\sin i = 7.5 \pm 4.0$ km s^{-1} (all errors herein are 1σ). Different spectral normalizations yield slightly different results but do not change the estimated mass of the B-type star (see below).

Although most of these values are in agreement with the results of Liu et al.[1], we note an effective temperature that is substantially lower than theirs, by almost 5,000 K. This is confirmed by an independent work[10] using non-LTE atmosphere models. Liu et al.[1] used a non-LTE grid[11] of atmosphere models that has a lower limit on $T_{\rm eff}$ of 15,000 K. This may have prevented their[1] ability to converge to temperatures below 15,000 K. Finally, using a Bayesian tool[12] to compare our measurements with stellar evolutionary models at solar metallicity[13], we obtain a best-fit evolutionary mass for the B-type star of $4.2^{+0.8}_{-0.7}M_\odot$.

Our results solve the apparent challenge posed by the presence of a very massive black hole at solar metallicity, in a mass range where pair-instability supernovae are expected with a very different end product. It also explains the absence of X-ray emission, which is normally generated by an accreting black hole. Further simulations of composite spectra formed by two approximately $4M_\odot$ main-sequence stars, where one of them is a rapidly rotating B-type star, show that if this latter star is rotating at a projected rotational velocity of more than about 200 km s^{-1}, it would be very difficult or impossible to detect given the quality of the data collected so far (see Supplementary Information). Alternatively, it is still possible that a black hole is present in LB-1; however, there is no observational evidence specifically requiring a very large black-hole mass (>50$M_\odot$).

A self-consistent picture of the intriguing LB-1 binary system may be found in a static Hα emission coming from a gaseous circumbinary disk similar to those seen in some B[e] stars[14] or from a more-massive counterpart of post-asymptotic giant branch phenomena[15]. Alternatively, the LB-1 system could be the progenitor of a low- to intermediate-mass X-ray binary, that is, a lower-mass B-type star in a binary system with a black hole.

Data availability

The data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

1. Liu, J. et al. A wide star–black-hole binary system from radial-velocity measurements. *Nature* **575**, 618–621 (2019).
2. Belczynski, K. et al. The formation of a 70-M$_\odot$ black hole at high metallicity. *Astrophys. J.* **890**, 113–859 (2020).
3. Groh, J. H. et al. Massive black holes regulated by luminous blue variable mass loss and magnetic fields – implications for the progenitor of LB-1. Preprint at http://arXiv.org/abs/1912.00994 (2019).
4. Ziółkowski, J. Evolutionary constraints on the masses of the components of the HDE 226868/Cyg X-1 binary system. *Mon. Not. R. Astron. Soc.* **358**, 851–859 (2005).
5. Casares, J. et al. A Be-type star with a black-hole companion. *Nature* **505**, 378–381 (2014).
6. Abbott, B. P. et al. GWTC-1: a gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs. *Phys. Rev. X* **9**, 031040 (2019).
7. Raskin, G. et al. HERMES: a high-resolution fibre-fed spectrograph for the Mercator telescope. *Astron. Astrophys.* **526**, A69 (2011).
8. El-Badry, K. & Quataert, E. Not so fast: LB-1 is unlikely to contain a 70 M$_\odot$ black hole. *Mon. Not. R. Astron. Soc.* **493**, L22–L27 (2020).
9. Tkachenko, A. Grid search in stellar parameters: a software for spectrum analysis of single stars and binary systems. *Astron. Astrophys.* **581**, A129 (2015).
10. Simón-Díaz, S. et al. A detailed non-LTE analysis of LB-1: revised parameters and surface abundances. *Astron. Astrophys.* **634**, L7 (2020).

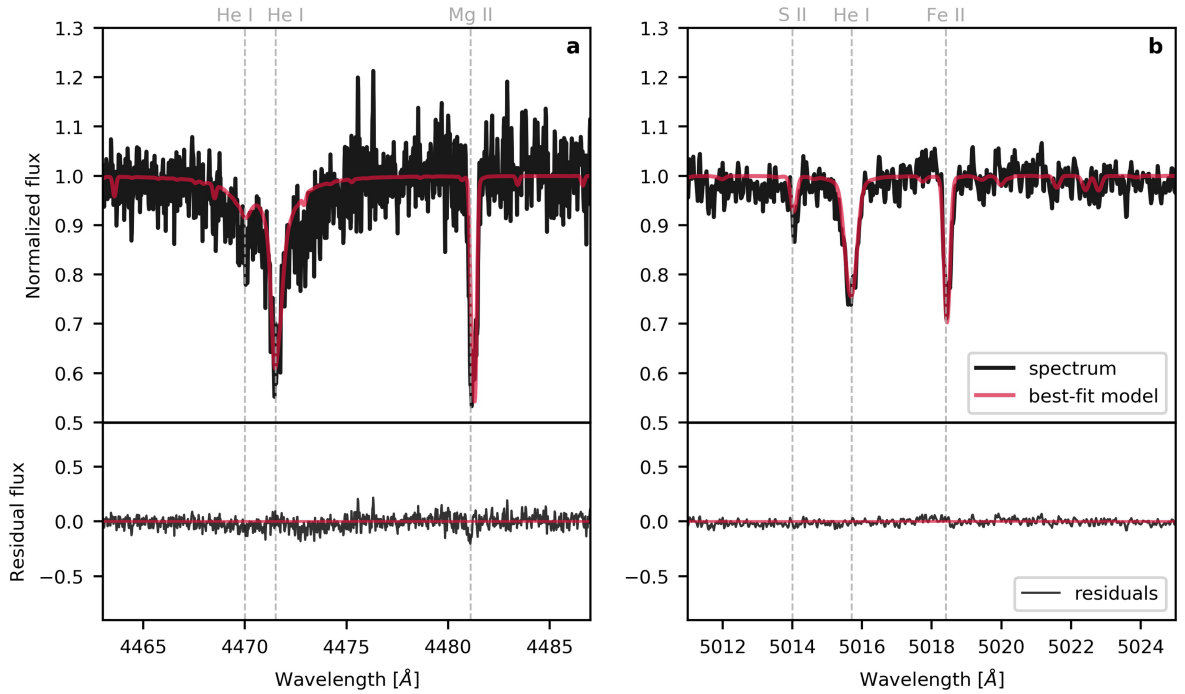
11. Lanz, T. & Hubeny, I. A grid of non-LTE line-blanketed model atmospheres of O-type stars. *Astrophys. J. Suppl. Ser.* **146**, 417–441 (2003).
12. Schneider, F. R. N. et al. Bonnsai: a Bayesian tool for comparing stars with stellar evolution models. *Astron. Astrophys.* **570**, A66 (2014).
13. Brott, I. et al. Rotating massive main-sequence stars. I. Grids of evolutionary models and isochrones. *Astron. Astrophys.* **530**, A115 (2011).
14. Condori, C. A. H. et al. The study of unclassified B[e] stars and candidates in the Galaxy and Magellanic Clouds. *Mon. Not. R. Astron. Soc.* **488**, 1090–1110 (2019).
15. Kamath, D. et al. Optically visible post-AGB stars, post-RGB stars and young stellar objects in the Large Magellanic Cloud. *Mon. Not. R. Astron. Soc.* **454**, 1468–1502 (2015).

Acknowledgements We acknowledge support from the Fonds Wetenschappelijk Onderzoek (FWO, Research Foundation Flanders) under project IDs G0F8H6N, G0B3818N, 12ZY520N, G0H5416N and GST-D6267-I002519N, and from the Onderzoeksraad (Research Council), KU Leuven under project IDs C16/17/007, C16/18/005 and C14/17/082. This project received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement numbers 772225 MULTIPLES and 670519 MAMSIE).

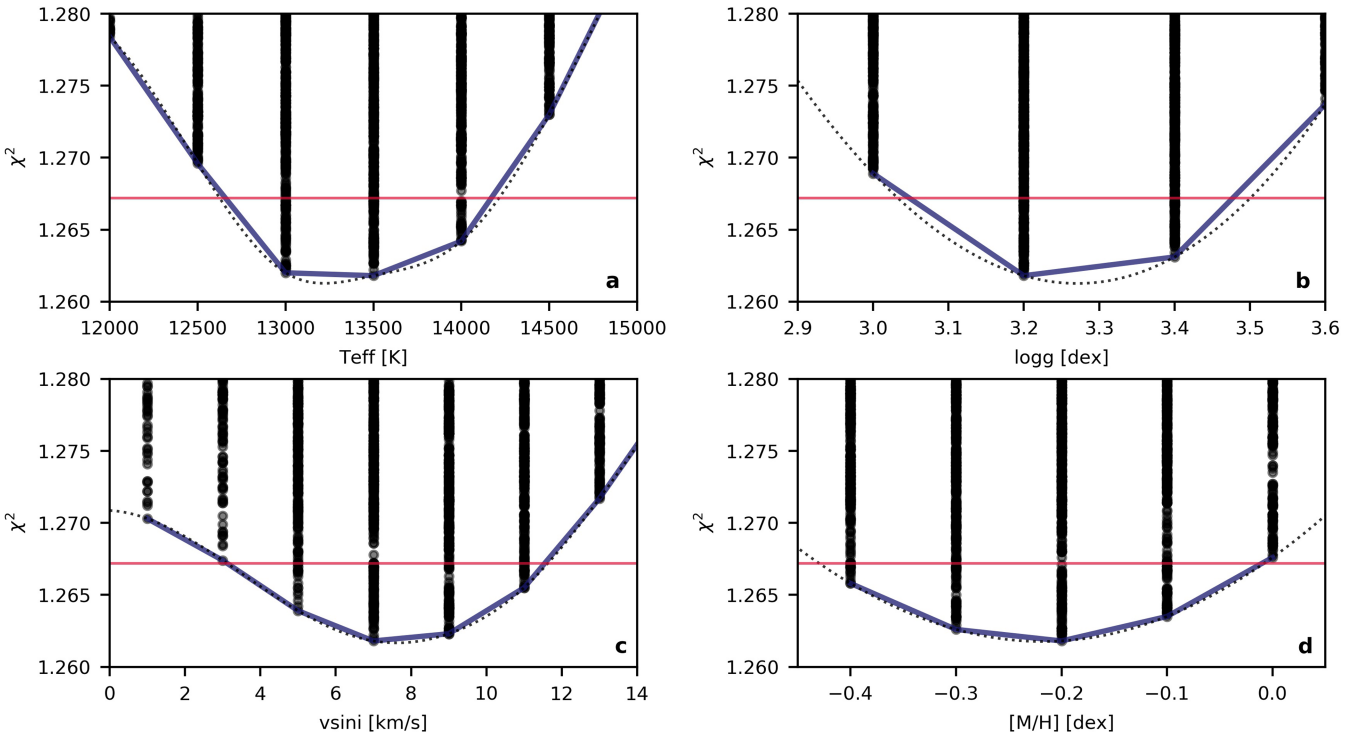
Author contributions This paper is based on an original idea by M.A.-M. J.B. prepared the observations. L.V. was the observer. H.V.W. and G.R. performed the HERMES data reduction. J.B. and L.M. performed the atmospheric analysis. M.A.-M., K.D., G.B. and T.S. studied the impact of the mixed profile on the radial-velocity measurements. D.M.B., E.B., M.F., L.M., T.S., A.T. and H.S. contributed to various aspects of the data analysis. All authors contributed to the discussion and interpretation of the results and commented on the written draft of the paper.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2216-x.
Correspondence and requests for materials should be addressed to H.S.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Stellar atmosphere fitting to the LB-1 optical spectrum. a, **b**, Top, best atmosphere model adjusted using the Grid Search in Stellar Parameters (GSSP) fitting suite (red) overlaid on the observed HERMES spectrum (black); and bottom, residuals of the fit. **a** and **b** each show a different small portion of the spectrum around the diagnostic lines of interest.



Extended Data Fig. 2 | Chi-squared distributions of the atmosphere fitting. **a**–**d**, Chi-squared maps of the GSSP atmosphere fitting projected on the one-dimensional parameter space for effective temperature (**a**), surface gravity (**b**), projected rotational velocity (**c**) and metallicity (**d**). The black data points are models, the blue solid and black dotted lines show linear and cubic interpolations through the best-fit grid points, respectively, and the horizontal red line corresponds to the adopted 1σ confidence threshold.

Reply to: On the signature of a 70-solar-mass black hole in LB-1

https://doi.org/10.1038/s41586-020-2217-9

Published online: 29 April 2020

Jifeng Liu[1,2,3]✉, Roberto Soria[2,4], Zheng Zheng[5], Haotong Zhang[1], Youjun Lu[1,2], Song Wang[1] & Hailong Yuan[1]

REPLYING TO: M. Abdul-Masih et al. *Nature* https://doi.org/10.1038/s41586-020-2217-9 (2020)

We reported[1] a wide black-hole binary LB-1, discovered through a designated radial-velocity-monitoring campaign of a large sample of stars. Our estimate of the mass of the black hole of approximately $70M_{\odot}$ rests on the following arguments, some of which have been questioned in the accompanying Comment by Abdul-Masih et al.[2] and by others[3–6]: (1) the Hα emission line comes from a disk around the black hole, seen moderately face-on as evidenced by its wine-bottle shape; (2) the barycentre motion of the Hα emission line wings traces the motion of the black hole and the absorption lines trace the motion of the donor star; and (3) the donor star has an effective temperature of $T_{\rm eff}$ = 18,100 ± 820 K, the surface gravity is log$g$ = 3.43 ± 0.15, and it has an evolutionary mass of $8.2^{+0.9}_{-1.2}M_{\odot}$. Here we re-examine those arguments, in light of more recent spectroscopic observations and further analysis. We show that the data still favour a very high mass of $23M_{\odot}$–$65M_{\odot}$ (with a most probable donor mass range of $5M_{\odot}$–$8M_{\odot}$); other lower-mass solutions are possible but less likely. Distinguishing definitively between solutions will require further Gaia astrometry data.

The key to the inference of an extraordinarily high black-hole mass in LB-1 is whether Hα and the other emission lines trace the motion of the unseen primary. Both Abdul-Masih et al.[2] and El-Badry et al.[3] demonstrate, correctly, that the Hα absorption line of the B-type star can induce an apparent barycentre motion of the Hα emission line wings that is similar to what we measured[1]. Their further inference of a static Hα emission line and its circumbinary origin, however, is unconvincing for several reasons.

First, it is too soon to claim a static Hα emission line on the basis of only one emission component and one absorption component, given the complexity of the profile. Hα may include contributions from: (1) emission, from the black-hole accretion disk, including the multi-scattering component that shapes the wine-bottle profile, as seen in nearly face-on disks; (2) emission, from the circumbinary disk; (3) absorption, from the atmosphere of the B-type star; (4) emission, from the stellar wind of the B-type star; and (5) emission, from hotspots or parts of the disk more strongly illuminated by the B-type star. In particular, irradiation from the B-type star causes alternating enhancement of the blue and red sides of the emission line, in phase with the orbital motion of the star. This leads to an apparent decrease of the observed barycentre motion of the emission line wings; that is, an effect opposite to that induced by the moving stellar absorption component. The likely presence of additional Hα emission from a moderate stellar wind[7] should also partly offset the absorption wings from the stellar atmosphere.

Second, neither the width nor the shape of the Hα emission line supports a circumbinary origin. Emission lines that are primarily from a circumbinary disk usually exhibit double-horn shapes with sharp edges and little extended wings (as observed in cataclysmic variable stars and B[e] stars)[8]. The edge velocity corresponds to roughly the projected Keplerian velocity at the inner radius of the disk, and would range from around 40 km s^{-1} for a $70M_{\odot}$ black hole to around 80 km s^{-1} for a black hole of the same mass as the B-type star. The observed Hα emission line shows a full-width at half-maximum of 220 km s^{-1} with substantial wings extended to ±500 km s^{-1}, clear evidence that its primary origin is not a circumbinary disk. Furthermore, the wine-bottle line profile is distinctly different from the double-horn shape of a circumbinary disk line, and can be attributed to multiple scatterings in the outer layers of the disk, as also pointed out by El-Badry et al.[3], provided that the disk is seen at a low inclination.

Nonetheless, we accept that the interpretation of the Hα profile is more complex than originally envisaged. A Doppler tomography investigation in progress may resolve the individual components and determine their kinematics. We found that we can more easily measure the orbital motion of the black hole by monitoring the optically thin Paschen emission lines, which have a cleaner and simpler double-peaked profile than the optically thick Hα line. We will present the results of our phase-resolved spectroscopic study of the near-infrared double-peaked Paschen emission lines in Liu et al. (manuscript in preparation). The main preliminary result of that study is that the peak position in the Pa β line clearly shows orbital motion in anti-phase with the B-type star, re-confirming the black-hole-disk origin of the Pa β line. The velocity amplitude is small, unsurprisingly, and suggests a black-hole-to-B-star mass ratio of 4.6–8.1 (preliminary), consistent with the previous estimate using Hα emission line wings.

The second issue of contention is the mass of the donor star, which, combined with the kinematic mass ratio, gives us the mass of the primary. Abdul-Masih et al.[2] obtained $T_{\rm eff}$ = 13,500 ± 700 K, log$g$ = 3.3 ± 0.3, and an evolutionary mass of $4.7^{+0.8}_{-0.7}M_{\odot}$. Simón-Díaz et al.[6] obtained $T_{\rm eff}$ = 14,000 ± 700 K, log$g$ = 3.5 ± 0.15, a spectroscopic mass of $3.2^{+2.1}_{-1.9}M_{\odot}$ (if the system is at the Gaia DR2[9] distance), and an evolutionary mass of $5.2^{+0.3}_{-0.6}M_{\odot}$. The mismatch between the spectroscopic mass and the evolutionary mass arises from different distance values and suggests that the single-star solution in Gaia DR2 may underestimate the true distance of LB-1 from Earth. The difference between these and our results can be attributed to the different datasets used for spectral modelling, and the different stellar atmosphere and evolution models adopted by the various groups. For example, if the PARSEC stellar models are adopted, the $T_{\rm eff}$ and log$g$ parameters in Abdul-Masih et al.[2] would instead correspond to $5.6^{+1.3}_{-1.2}M_{\odot}$, as shown in figure 2 of Simón-Díaz et al.[6]. Even if the mass of the donor star is indeed approximately $5M_{\odot}$ instead of

[1]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China. [2]School of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]WHU-NAOC Joint Center for Astronomy, Wuhan University, Wuhan, China. [4]Sydney Institute for Astronomy, School of Physics, The University of Sydney, Sydney, New South Wales, Australia. [5]Department of Physics and Astronomy, University of Utah, Salt Lake City, UT, USA. ✉e-mail: jfliu@bao.ac.cn

approximately $8M_{\odot}$, the black-hole mass would be about $23–41M_{\odot}$, for a black-hole-to-B-star mass ratio of 4.6–8.1.

There is another possible class for the donor—a stripped helium star in a short-lived evolutionary phase[4]. Irrgang et al.[5] pointed out possible abundance patterns in the spectra for such a stripped star. From their best-fit values of $T_{\rm eff} = 12{,}720 \pm 260$ K and $\log g = 3.00 \pm 0.08$, they derived a spectroscopic mass of $1.1 \pm 0.5M_{\odot}$. However, Abdul-Masih et al.[2] argued that there is no observational evidence for such a stripped star. Another challenge to this scenario is the short lifetime of these stars, and thus the low probability of discovering one among the 3,000 stars in the LAMOST survey[3].

Although it is difficult to determine the donor mass unambiguously from spectroscopy alone, Gaia astrometry can ultimately solve the problem. As we pointed out[1], Gaia transit data can reach an astrometric error as small as 0.1 mas per visit, enough to resolve the binary wobble of LB-1. The full orbit and the parallax can be solved simultaneously by combining radial-velocity measurements and Gaia transit data. This will give the total mass of the binary by Kepler's third law, and hence the donor mass and the black-hole mass from their mass ratio.

1. Liu, J. et al. A wide star–black-hole binary system from radial-velocity measurements. *Nature* **575**, 618–621 (2019).
2. Abdul-Masih, M. et al. On the signature of a 70 solar-mass black hole in LB-1. *Nature* https://doi.org/10.1038/s41586-020-2216-x (2019).
3. El-Badry, K. & Quataert, E. Not so fast: LB-1 is unlikely to contain a 70 $M_{\odot}$ black hole. *Mon. Not. R. Astron. Soc.* **493**, L22–L27 (2020).
4. Eldridge, J. J. et al. Weighing in on black hole binaries with BPASS: LB-1 does not contain a $70M_{\odot}$ black hole 2019. Preprint at http://arXiv.org/abs/1912.03599 (2019).
5. Irrgang, A. et al. A stripped helium star in the potential black hole binary LB-1. *Astron. Astrophys.* **633**, L5 (2020).
6. Simón-Díaz, S. et al. A detailed non-LTE analysis of LB-1: revised parameters and surface abundances. *Astron. Astrophys.* **634**, L7 (2020).
7. Rosendhal, J. D. A survey of H-alpha emission in early-type high-luminosity stars. *Astrophys. J.* **186**, 909–937 (1973).
8. Condori, C. A. H. et al. The study of unclassified B[e] stars and candidates in the Galaxy and Magellanic Clouds. *Mon. Not. R. Astron. Soc.* **488**, 1090–1110 (2019).
9. Gaia Collaboration. Gaia Data Release 2. Summary of the contents and survey properties. *Astron. Astrophys.* **616**, A1 (2018).

Acknowledgements We thank S. Justham, W. Hamann, S. Wang and many others for discussions. This work was supported by the National Science Foundation of China (NSFC) under grant numbers 11988101/11933004 (J.L.), 11690024 (Y.L.) and 11603035 (S.W.). This work was also supported by the National Key Research and Development Program of China (NKRDPC) under grant numbers 2019YFA0405504/2016YFA0400804 (J.L.) and 2016YFA0400704 (Y.L). J.L. acknowledges support by the Tencent Xplore Prize.

Author contributions The present author list includes only those authors of the original paper who have now contributed substantially to the writing of this Reply. J.L. and R.S. drafted the manuscript, and all other authors contributed substantially to the discussion and revision.

Competing interests The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.L.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author Correction: Temporal plasticity of apical progenitors in the developing mouse neocortex

https://doi.org/10.1038/s41586-020-2218-8

Correction to: *Nature* https://doi.org/10.1038/s41586-019-1515-6

Published online 28 August 2019


Check for updates

Polina Oberst, Sabine Fièvre, Natalia Baumann, Cristina Concetti, Giorgia Bartolini & Denis Jabaudon

In this Article, some of the histological sections displayed in Extended Data Figs. 2b and 7 were mistakenly imaged twice and considered as distinct data points (two image pairs in Extended Data Fig. 2b and five image pairs in Extended Data Fig. 7, highlighted by coloured boxes in Fig. 1 of this Amendment). This error happened because the images in question were taken several days or weeks apart by two distinct investigators; differences in image acquisition and processing led to different image properties, resulting in their misclassification. Returning to the original sections and images, we have now corrected this error by removing one copy of the images that were acquired twice and ascribing the corresponding neurons to their proper oligoclones (see Fig. 1 of this Amendment). As a result, 2 cells (out of 403) in the $AP_{12\rightarrow12}$ condition and 11 cells (out of 716) in the $AP_{12\rightarrow15}$ condition have been removed from the dataset. The significance of all statistical analyses following this correction remains unchanged (see the statistical table and summary of figure changes in the Supplementary Information of this Amendment). The Supplementary Table of this Amendment describes the changes to the original Supplementary Table, in which affected values are highlighted using the same colour code as in Fig. 1. In sections of the original Methods, some values have changed slightly (the changes to the affected paragraphs are detailed in the Supplementary Information to this Amendment). In addition, the legend of Extended Data Fig. 3b now cites Extended Data Fig. 2 instead of Extended Data Fig. 4 and the legend of Extended Data Fig. 6 now mentions 79 oligoclones ($AP_{12\rightarrow15}$) instead of 78. The original Supplementary Table 1 and Extended Data Figs. 2b and 7 of the original Article have been corrected online, in addition to the text changes described above (but not the original figures).

Supplementary Information is available in the online version of this Amendment.

Author Correction: *Listeria monocytogenes* impairs SUMOylation for efficient infection

https://doi.org/10.1038/s41586-020-2154-7

Correction to: *Nature* https://doi.org/10.1038/nature08963

Published online 22 April 2010.

David Ribet, Mélanie Hamon, Edith Gouin, Marie-Anne Nahori, Francis Impens, Hélène Neyret-Kahn, Kris Gevaert, Joël Vandekerckhove, Anne Dejean & Pascale Cossart

In Fig. 4d of this Letter, the band corresponding to free SUMO1 in the left-hand blot is incorrect, as it is a repeat of the band corresponding to free SUMO2 in the right-hand blot. Figure 1 of this Amendment shows the original incorrect panel and the corrected panel of Fig. 4d, for transparency to readers. The full uncropped blots are shown in the Supplementary Information to this Amendment. The original Letter has not been corrected.

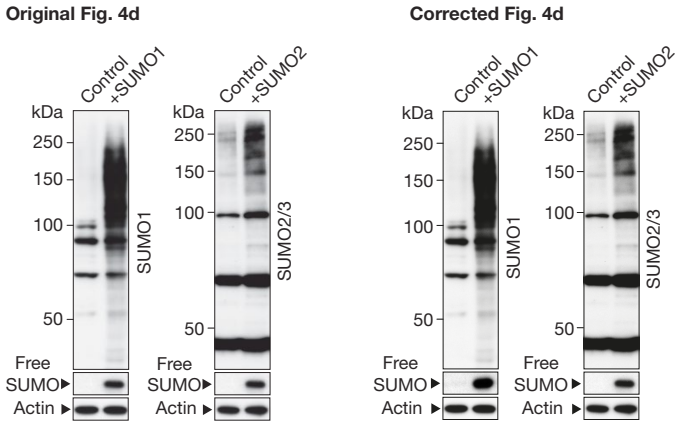


Fig. 1 | This figure displays the corrected and the incorrect published Fig. 4d of the original Letter.

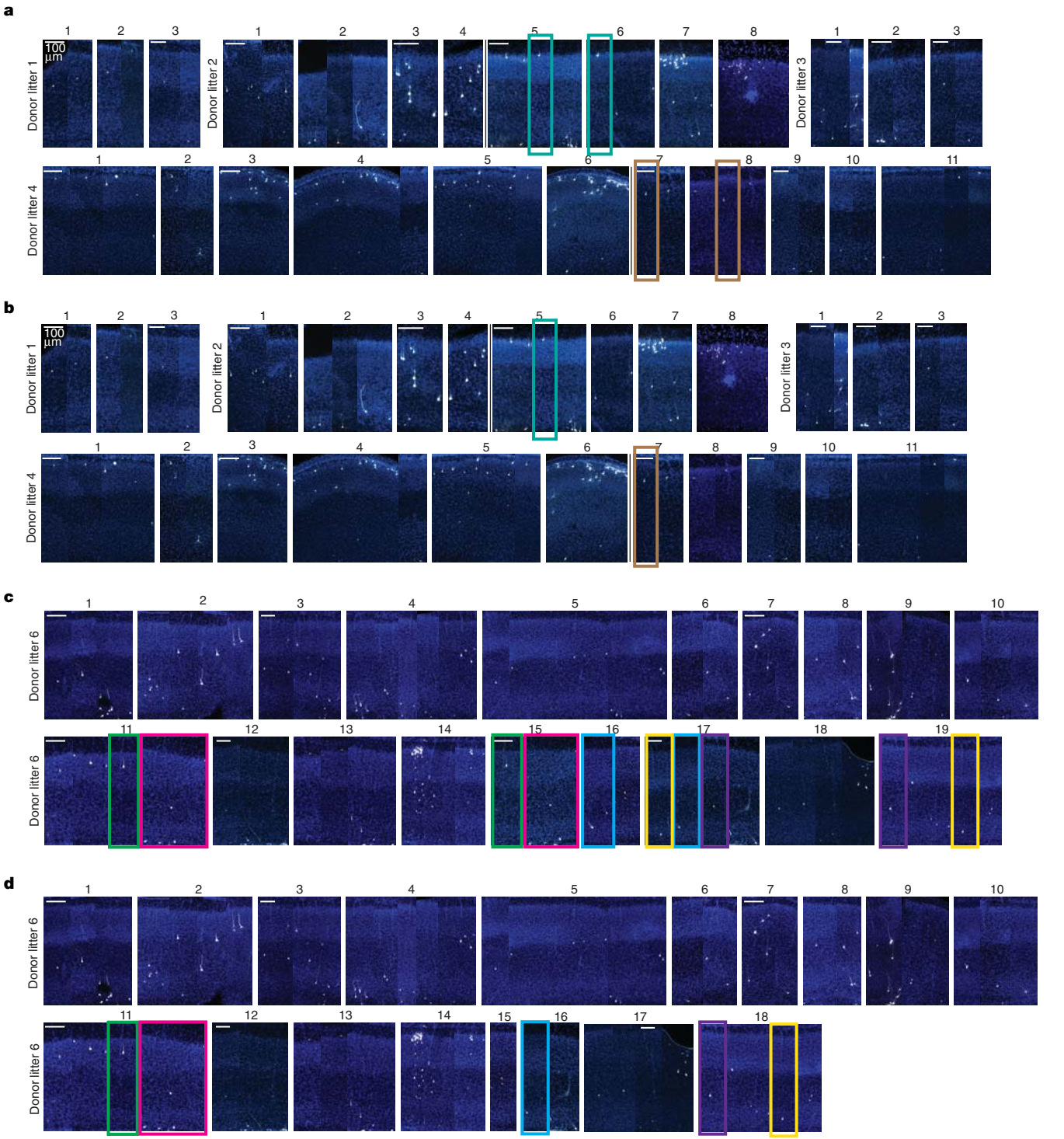


Fig. 1 | This figure shows the as-published, incorrect panels of Extended Data Figs. 2b and 7 and the corresponding corrected panels. a, **b**, In this part of Extended Data Fig. 2b, the sections that were imaged twice (in 'Donor litter 2' and in 'Donor litter 4') are boxed in blue and brown, respectively (**a**) and the retained images are boxed (**b**). **c**, **d**, In this part of Extended Data Fig. 7, the five sections that were imaged twice (in 'Donor litter 6') are boxed in matching colours (**c**) and the retained images are boxed (**d**).

nature

index

Annual tables

SHOWING THE WAY

Our 50 world leaders in natural-sciences research

nature
index

Annual tables

Editorial Catherine Armitage, Bec Crew, Rebecca Dargie, Gemma Conroy, David Payne **Analysis** Bo Wu, Catherine Cheung **Art & design** Madeline Hutchinson, Tanner Maxwell, Wojtek Urbanek **Production** Jason Rayment, Ian Pope, Nick Bruni, Bob Edenbach, Joern Ishikawa **Marketing & PR** Stacy Best Ruel, Angelica Sarne, Elizabeth Hawkins **Sales & partner content** Sabrina Ma, Pinky Zhang, Yingying Zhou, Ruffi Lu **Publishing** Rebecca Jones, Richard Hughes, David Swinbanks

Nature Index 2020 Annual Tables a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. This publication is based on data from the Nature Index, a Nature Research database, with a website maintained and made freely available at natureindex.com.

Nature editorial offices
The Campus, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

Customer services
To advertise with the Nature Index, please visit natureindex.com or email clientservicesfeedback@nature.com.

As clinicians, world leaders and policymakers race to respond to the coronavirus pandemic, publication speed and transparency have never been more important. Research will inform how we overcome these unprecedented challenges.

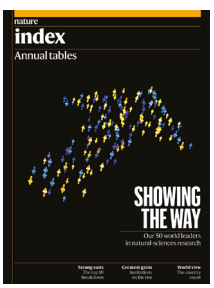
With the release of the *Nature Index 2020 Annual Tables*, we celebrate the institutions and countries producing high-quality natural-sciences research. Our metric, Share, formally referred to as Fractional Count (FC), is based on an institution's or country's contribution to articles published in 82 journals tracked by the Nature Index database. These journals were selected by committees of 58 leading researchers in the natural sciences, who were asked to nominate the journals in which they would most like to publish their best work. Their deliberations were validated by a survey of more than 6,000 scientists worldwide.

Although the Nature Index Annual Tables compare institutions and countries based on counts of their research outputs in Nature Index journals, our measures alone do not tell the whole story. We recognize, in line with the San Francisco Declaration on Research Assessment, that outputs from scientific research include not only journal articles, but data, software, intellectual property and highly trained young scientists. We encourage readers to weigh Nature Index data alongside information from other sources when considering research quality and institutional performance.

In the *Nature Index 2020 Annual Tables*, we see strong performances from institutions with considerable funding and reputation, such as the Chinese Academy of Sciences, Harvard University and the Max Planck Society. There are surprises, too, as the University of Science and Technology of China (USTC) enters the top ten in 8th position, rising from 17th in the 2019 Annual Tables, with a 25.58% increase in adjusted Share* (see S44) in the Nature Index. In the Rising Stars table, which tracks an institution's growth in output from 2015 to 2019, USTC is second only to the University of Chinese Academy of Sciences, which achieved a 242% increase in adjusted Share.

For stories of achievement that go beyond the metrics, as well as Rising Stars tables in the life sciences, physical sciences, chemistry, and Earth and environmental sciences, visit www.natureindex.com.

Bec Crew
Senior Editor
Nature Index

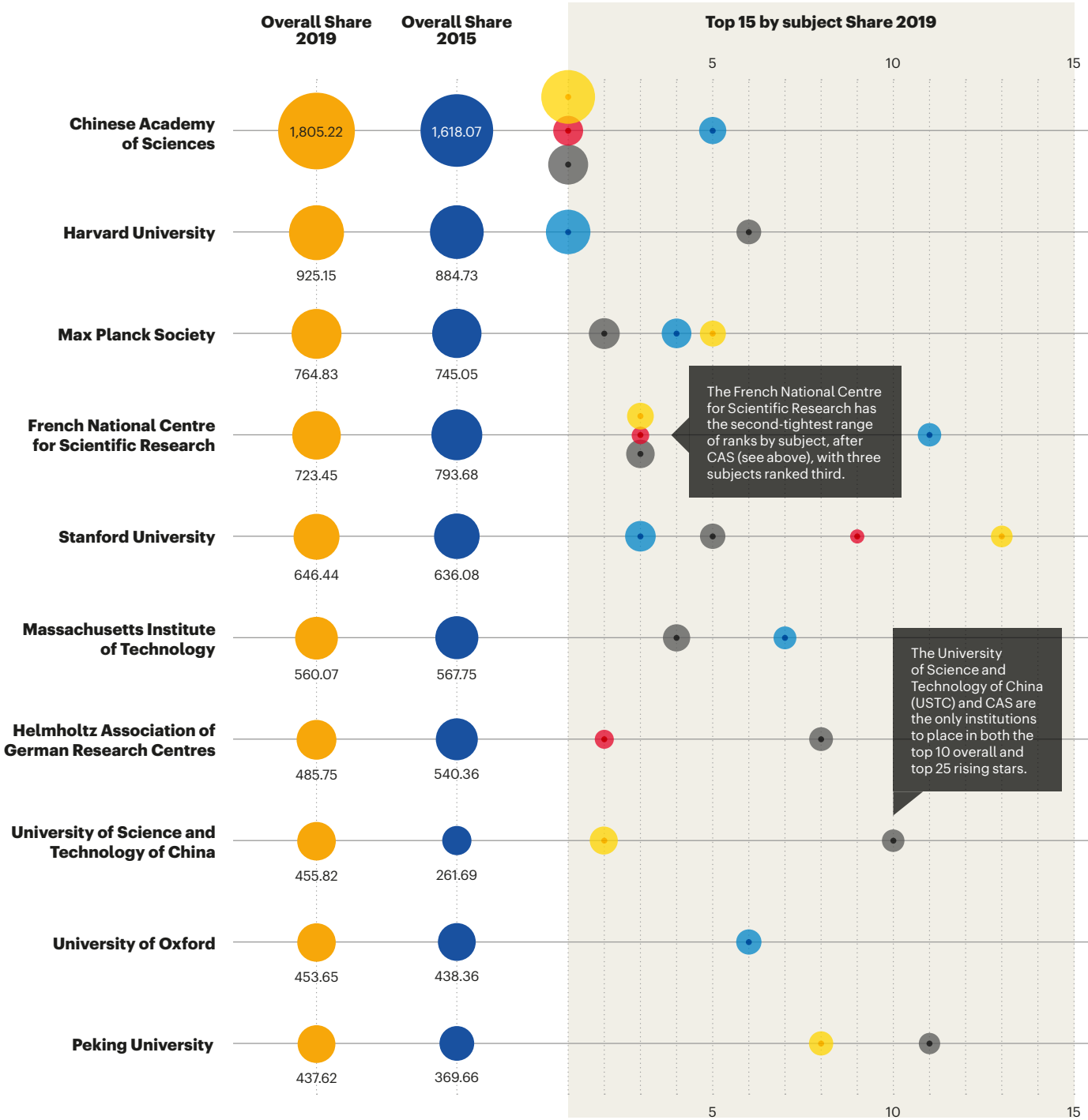


Cover design:
Tanner Maxwell

Contents

Show of strength

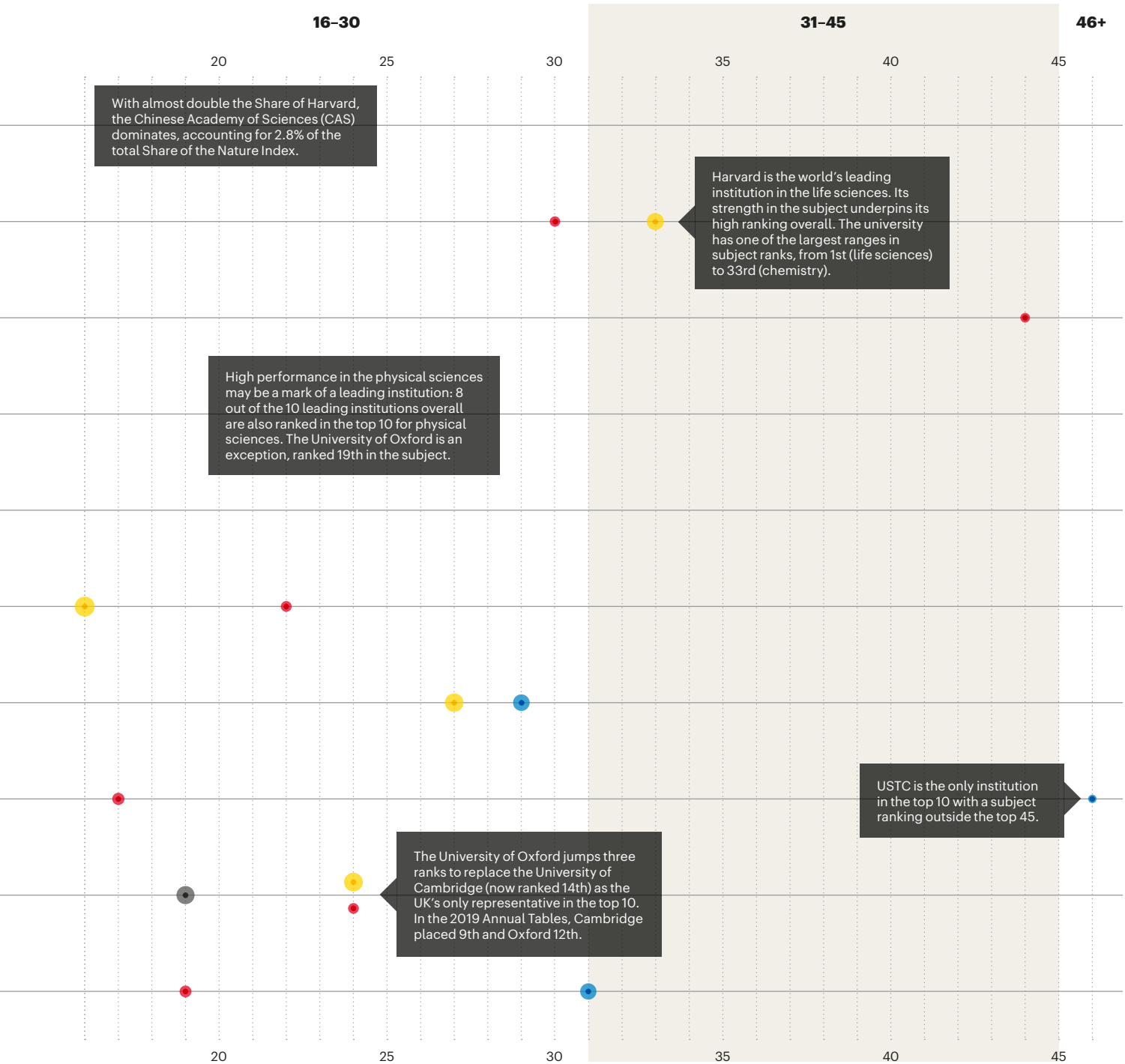
These institutions lead in natural-sciences research in journals tracked by the Index.
Data analysis by Catherine Cheung. Infographic by Tanner Maxwell.



HOW TO READ THIS GRAPH

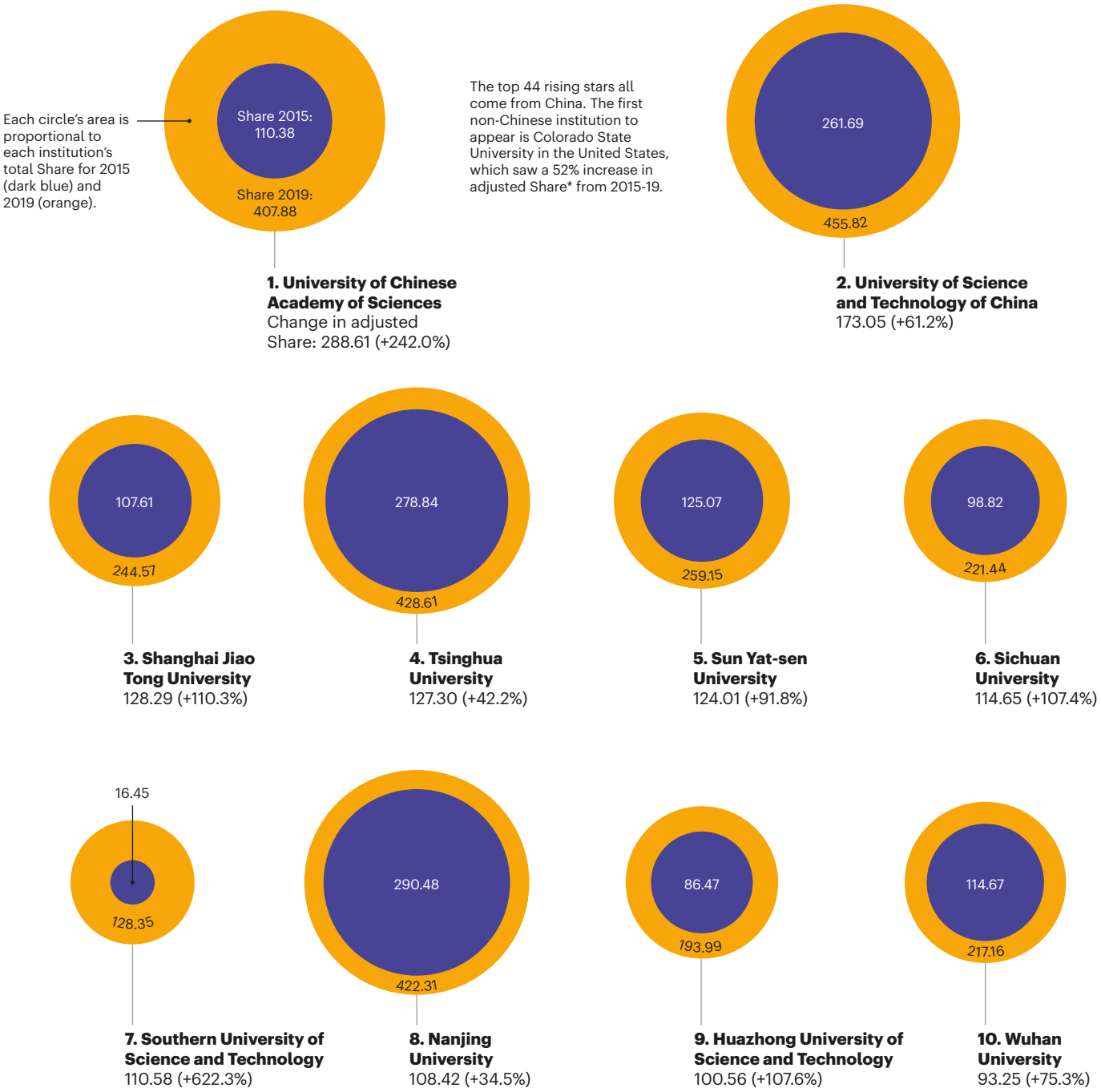
The top 10 global institutions in the *Nature Index 2020 Annual Tables* are listed here with circles representing each institution's overall Share in 2019. Overall Share in 2015 is shown for comparison. Each circle is coloured according to year or subject, and each circle's area is sized proportionally to Share. Subject circles are charted across both pages according to each institution's rank.

- Overall Share 2015
- Overall Share 2019
- Chemistry
- Earth and environmental sciences
- Life sciences
- Physical sciences



Fastest movers

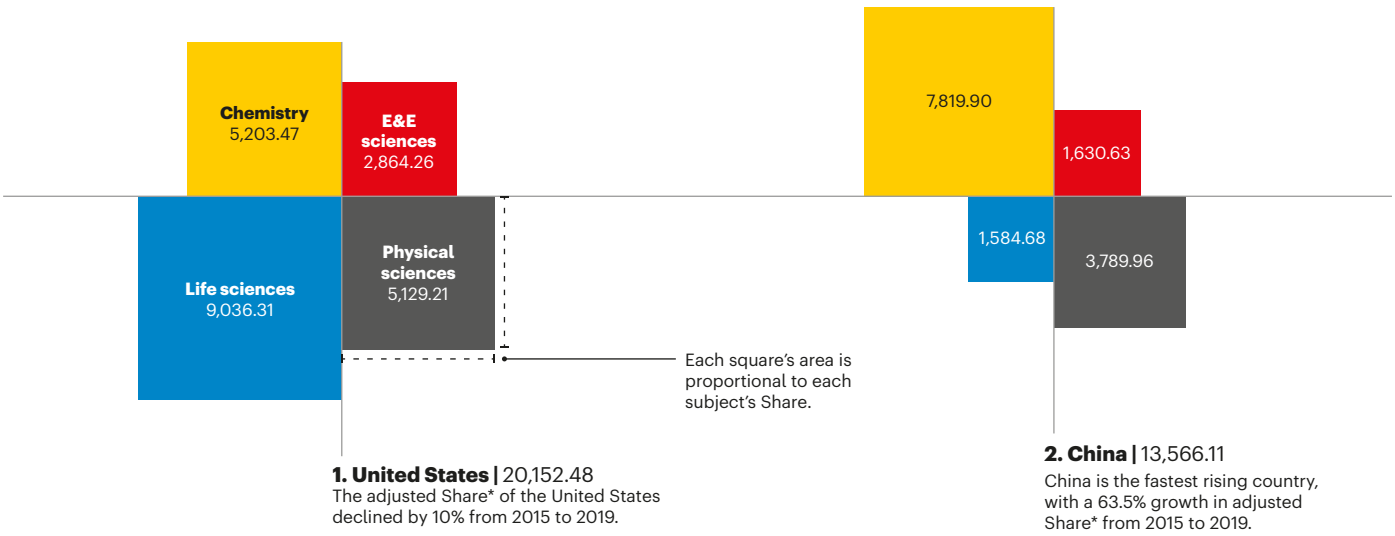
China dominates the rising stars ranks, ordered by change in adjusted Share.
Data analysis by Catherine Cheung. Infographic by Tanner Maxwell.



* When comparing data over time, Share values are adjusted to 2019 levels to account for the small annual variation in the total number of articles in the Nature Index journals.

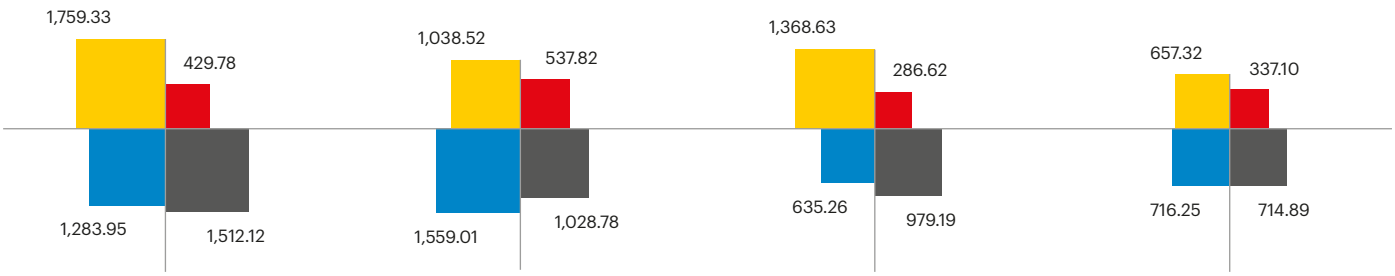
A world of progress

Subject strengths reveal high-performing hubs in all natural-sciences discipline areas. **Data analysis by Catherine Cheung. Infographic by Tanner Maxwell.**



1. United States | 20,152.48
The adjusted Share* of the United States declined by 10% from 2015 to 2019.

2. China | 13,566.11
China is the fastest rising country, with a 63.5% growth in adjusted Share* from 2015 to 2019.



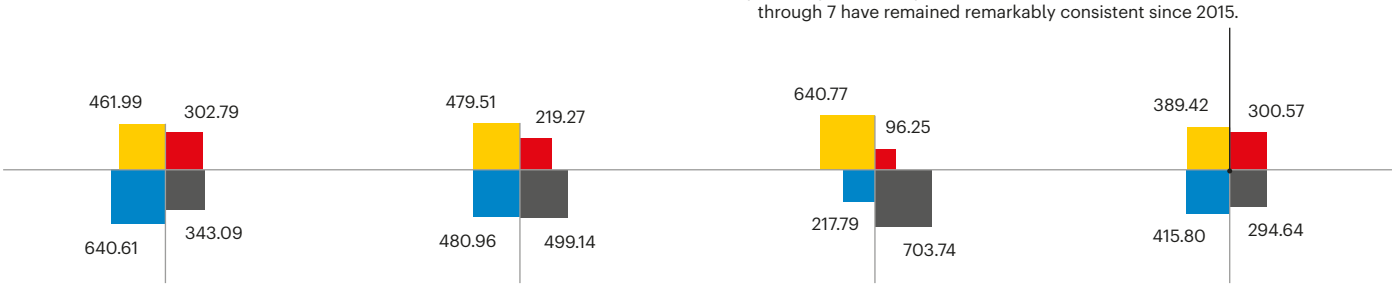
3. Germany
4,545.70

4. United Kingdom
3,773.66

5. Japan
3,024.32

6. France
2,238.55

Australia has held 10th place for two years' running, having displaced Spain to 11th place in the 2019 Annual Tables. Ranks 1 through 7 have remained remarkably consistent since 2015.



7. Canada
1,602.09

8. Switzerland
1,487.88

9. South Korea
1,435.23

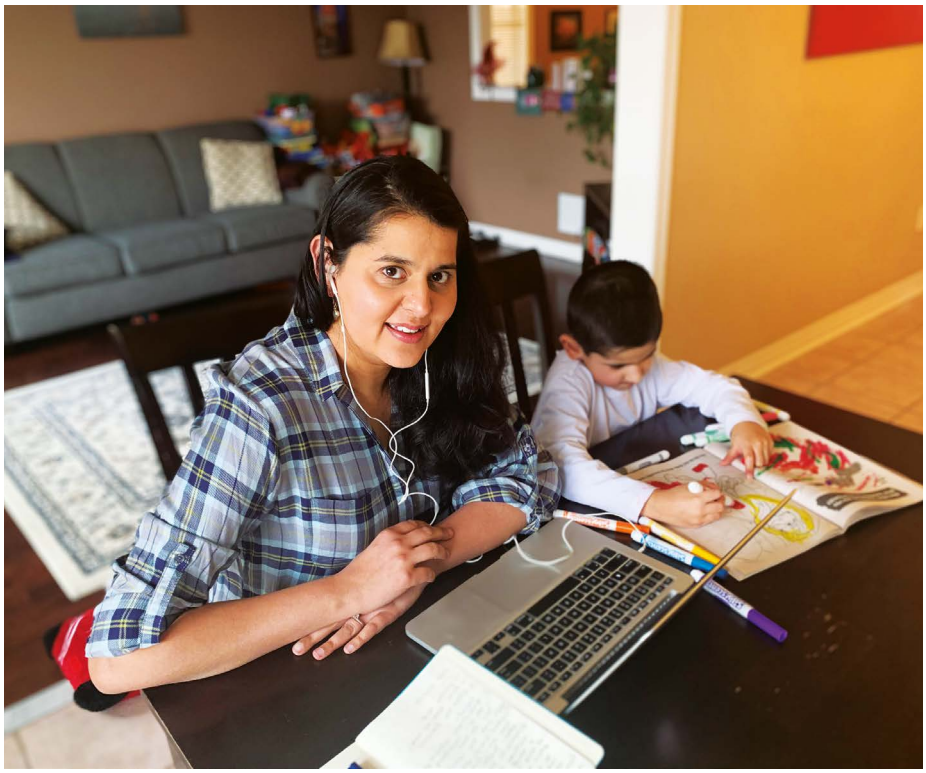
10. Australia
1,259.95

* When comparing data over time, Share values are adjusted to 2019 levels to account for the small annual variation in the total number of articles in the Nature Index journals.

Work

Your story

Send your careers story to: naturecareerseditor@nature.com



SAPNA SHARMA

Ecologist Sapna Sharma working at home in Canada with her son.

FROM LAB BENCH TO LEGO BRICKS

Scientist-parents describe their efforts to juggle family and work duties during virus restrictions.

Scientists who have children are trying to manage their productivity as employers, universities and schools worldwide have closed in an effort to contain the COVID-19 pandemic. Here's what six researchers are doing to navigate the tensions that arise when full-time work and full-time parenting intersect at home.

SAPNA SHARMA
WORK IN BURSTS

Our public schools closed on 13 March. I have enjoyed the extra time with my son, who is four years old, and whom we are home-schooling for several hours a day. But I have had to move meetings online through the video-conferencing service Zoom. I hold these approximately one-hour laboratory and collaboration meetings while my son is awake, during regular working hours. He usually does an independent activity then, such as colouring, playing with his toy cars and trains or working on his sticker collection, or he watches a cartoon. Sometimes, he comes to sit on my lap and joins the Zoom meetings. I am fortunate to work with colleagues who are patient and understanding, and enjoy my son popping in.

I also work in shorter stints — 45–60 minutes at a time during the day — while my son is busy with his own activities in the same room. During this time, I can do only teaching, service, editing and outreach work. It's difficult to write or work on research during this time, so I try to do that before my son wakes up at 9:30 a.m. and after he has gone to bed. Depending on the day, and because it was especially busy while transitioning my teaching to online courses, I work a total of 6–8 hours a day. My research, which I can do remotely, involves analysing large data sets to understand the impacts of climate change on lakes.

I make detailed notes on what I absolutely need to do when I have time for work, and go straight to my to-do list as soon as I get a chance. I am fortunate to have an incredibly supportive husband, who helps to give me focused work time when I need it. He has a PhD in particle physics and works in quantitative-risk management, and is also working at home during this pandemic. He and I alternate caring for our son, typically for a couple of hours at a time. For example, if I am in a meeting or teaching, my husband will play with and take care of our son during that time, and vice versa. If we are both in a meeting or working at the same time, our son will watch something on television.

Our schedule of meetings with others is pretty well structured. Work time is not structured, and we found that we're both putting in very long days to accommodate working and taking care of our son.

When we are both busy or in a virtual class or meeting, our local friends and family help us out by 'playing' with our son over video call. My parents live outside Toronto and my husband's parents are in Chicago, Illinois. Both sets had offered to help in person during this time, but we asked them not to, because of the higher risk of infection for them.

I recognize that I am not going to be as productive during the COVID-19 outbreak, but I am comfortable with that. Academia is a marathon rather than a sprint, and I will have time to be productive after this public-health crisis is over.

Sapna Sharma is a freshwater ecologist at York University in Toronto, Canada.

Editor's note: Sharma e-mailed on 1 April with this update: "I became ill on 25 March with symptoms consistent with COVID-19 and with a kidney infection. I learnt on 1 April that I tested negative for the virus, and antibiotics are helping with the kidney infection. I taught my last class for the semester on 25 March, and have taken a break from work. I will work again when I'm feeling better." Sharma e-mailed again on 3 April to say that she was feeling better and was back to work. *Nature* encourages readers who are feeling ill to take sick leave.

JAVIER G. FERNANDEZ
NEGOTIATE WITH KIDS

My campus set a social-distancing policy on 17 February, shortly after the Singaporean government raised its alert level in response to COVID-19. All personnel were assigned to one of two teams that alternate between working on campus and working from home, to minimize contact between individuals. Most classes have now moved to an online format.

I'm much more productive because of the reduced administrative workload and the lack of 'important' meetings that, apparently, can be swapped for a couple of e-mails.

When I work from home now — if I have a phone or online meeting after 5 p.m., or if it was a day when our three-year-old daughter's school was closed — I always make people aware that she will be around and that the meeting might be interrupted. [Schools and workplaces in Singapore were open on certain days at the time of this interview.] My advice is, don't stress over trying to hide that your children are home with you. Being a person doesn't make you less of a professional.

I share all tasks related to childcare and the household with my spouse, who is a product manager for the online shopping platform Alibaba. In general, this task-sharing happens organically, but we have some basic rules. For example, I picked up our daughter from school every Monday, Wednesday and Thursday, and my spouse picked her up every Tuesday and Friday. So whoever was not collecting her could work longer hours on those days.

That is not entirely written in stone. We make changes occasionally, but communication is crucial for those specific arrangements.

If we are both working from home, we divide childcare into morning and afternoon sessions: I take care of our daughter in the morning while my spouse focuses entirely on work, and we swap in the afternoon. Sometimes, one of us might get some work done during our childcare session, but looking after our daughter must be the priority during those hours. Our system works for us because we perceive the process as a collaborative effort, rather than as a competition.

Being strict with time is crucial. Set precise schedules and focus on what you are doing. As we say in Spanish (I'm from northern Spain), '*No se puede estar en misa y repicando,*' which literally translates as, 'You can't be attending mass and ringing the bells at the same time.' We use it to exemplify two things that you might want to do at the same time, but which you can't do right simultaneously.

Strict schedules help our daughter to become used to a routine, and she contributes to it instead of fighting it. I negotiate deals with her. For example, if I need to focus for 1–1.5 hours, I ask her to play on her own during that time. She can stay engaged for longer periods by drawing, painting or playing with Duplo big blocks. I commit to my promise that if she does it, we will then play or read or go outside together.

Younger children might be able to focus for only 30 minutes at a time. Pick a time frame that is achievable for your child, so that you will both feel successful. The important thing in these negotiations is to be strict not only with the working hours, but also with the playing hours. In the end, if you embrace those, they become an excellent way to release stress.

It doesn't work all the time. Kids are kids, so there is inherent randomness in them, but they are growing and are incredibly good learners. So you can grow together.

Javier G. Fernandez is a materials scientist at Singapore University of Technology and Design in Singapore.

Researcher Javier Fernandez's baby daughter helps with paperwork at home in Singapore.

S. FABRE

MARICA BRANCHESI
SHARE THE DAY

The schools here in central-south Italy closed on 4 March. My children, who are three and four years old, my partner and I are all together in a 100-square-metre apartment on the first floor of our building. We are lucky to have a large balcony, where we spend a lot of time in sunny weather. Under a 9 March federal order, we can go out only for food or medicine. The order, which recommended as much remote working as possible, makes it impossible for me to meet with my students and postdocs.

By 16 March, our institute's entire staff was working remotely for an indefinite period. Institute buildings remain open for limited hours for document retrieval only.

I am separated from my large extended family — grandparents, parents, sister, uncles and cousins — who all live a three-hour drive away in Urbino, central Italy. There is the devastating awareness that, if something bad happens, we cannot meet or go to help each other. We have video or phone calls every day with my parents, my sister's family and with my grandparents, who are both more than 90 years old.

It is difficult to work and to concentrate, but it is important to think of something else and to continue with one's life. I am lucky because I can continue to work remotely.

Life goes by slowly, between the desire to work as if nothing special is happening, and the search for information on the coronavirus,

waiting for numbers that might indicate an improvement in the situation. I share the day with my partner, a physicist who also works at the Gran Sasso Science Institute. One of us works while the other spends time with the children, devising games to stop them from getting bored, and then we swap roles. Some days my partner is more busy with work; other days, I am. We often work for three or four hours after the children have gone to bed, sometimes as late as 2 a.m. while they sleep. It is not particularly efficient because I am usually tired by then.

The work has not changed much — data from telescopes, satellites and gravitational-wave detectors can still be received on our personal computers. However, I miss the human contact, and the ideas that arise during long group discussions using blackboards, or by sharing notes and printed documents in person. We continue to explore the Universe, but we feel a little more alone.

Marica Branchesi is an astrophysicist at Gran Sasso Science Institute in L'Aquila, Italy.

ANTHONY TRAN
AUTOMATE WORKFLOWS

When my 18-month-old son's day-care centre closed in late March, my wife and I had prepared for it. She is an actuary for a large health-care company and works from home, so we rotate childcare shifts on the basis of our availability. We schedule our tele-meetings so that they do not overlap, leaving at least one of us to be available to attend to our son if necessary.

Communication with managers is crucial. Be transparent about what will work for you as a result of your child being at home. In my experience, managers will generally be understanding and try their best to accommodate your schedule. Your family life is as important as business operations and should be treated as such, especially during a rare global crisis.

In San Mateo county in California, we have a 'shelter-in-place' order that requires us to stay at home except for essential activities such as buying food. This is a good time to use toys and activities with children that encourage more-focused thought and creativity, which in turn can buy you more time to get work done. Kids can perform more-elaborate activities that engage them for longer periods, so that parents have longer uninterrupted periods in which to be productive. When time to work is scarce in general, it can make a difference between a productive and an unproductive day. For example, we recently bought a magnetic tile set that helps to keep our boy entertained for longer periods. Inevitably, however, he gets bored with that, so we keep different sets of toys and books on rotation.

My work involves programming laboratory robots that process clinical samples for cancer diagnostics. Every week, there's some preliminary work that I can do at home for a few days. Then I spend a few half-days in the lab to test the processes out on actual systems. Our work falls into the 'health-care operations' category of essential businesses, which is exempt from San Mateo's shelter-in-place order.

Because the processes are automated, it minimizes the time I need to be in the lab, and I can go home for a while to care for my son. Colleagues who can be in the lab a bit more often set up and run most of the automated experiments. This means that teamwork, relationship-building and communication are especially important.

Anthony Tran is a process engineer at a health-care company in San Francisco, California.

Editor's note: Tran e-mailed these updates on 26 and 29 March: "Wife and I have been sick. Throw in a kid to take care of and we were barely scraping by. But our four-to-five-day fevers broke yesterday, so we're doing much better. Don't know if it was COVID-19, but it seems we're on the road to recovery."

HIROMI IINUMA
LET GO

My seven-year-old son's school closed on 5 March. My spouse is also a particle physicist, but his office is in a controlled-radiation area, so our son cannot stay with him during the day. Throughout the pandemic, my spouse has been doing his best at work and at home, but I have been spending more time with my son.

My son has asthma, so I hesitate to take him to day care, which remains open for now. Fortunately, I have my own university office, which I can still use, so I work there with my son. I would work from our two-bedroom house, but I need to have in-person discussions with my students so that I can guide them intensively and better assess how they are doing during this difficult time.

I have no idea how to maintain my research productivity while working alongside my son. My husband says that my attitude changes completely when I am doing my work. My son told me early on that he thought I hated him, because of my attitude when I worked in his presence. I never realized that I put such stress on him. It's a dilemma, trying to be a good mother and a researcher at the same time.

I give him many workbooks, origami, whatever he wants. But he can stay calm for only a few hours. My research productivity has diminished quantitatively and qualitatively. At the end of the day, both my son and I are very exhausted. Of course, I work during the weekends at home, too. At those times, my son can play with his father.

The Vidyo video-conferencing system is my best tool for remote discussions with colleagues. At times, we hold small meetings in person on campus, but we prefer to use the online system so we can each remain in our own offices and avoid non-essential contact. We have postponed many non-urgent meetings. I see now that many were not essential.

My colleagues accept my remote attendance at important meetings. And, fortunately, throughout March I did not have to visit an experimental area. But I am really wondering how long this situation will continue beyond April.

Hiromi Iinuma is a particle physicist at Ibaraki University in Mito, Japan.

SEYED AKBAR JAFARI
EXPLOIT THE BALCONY

I have a seven-year-old daughter and a six-month-old son. My wife is an IT manager and has worked from home, too, since early March. My daughter is not a challenge, because we have only to take care of her online homework. The challenge is my son. I have roughly divided these caring duties with my wife.

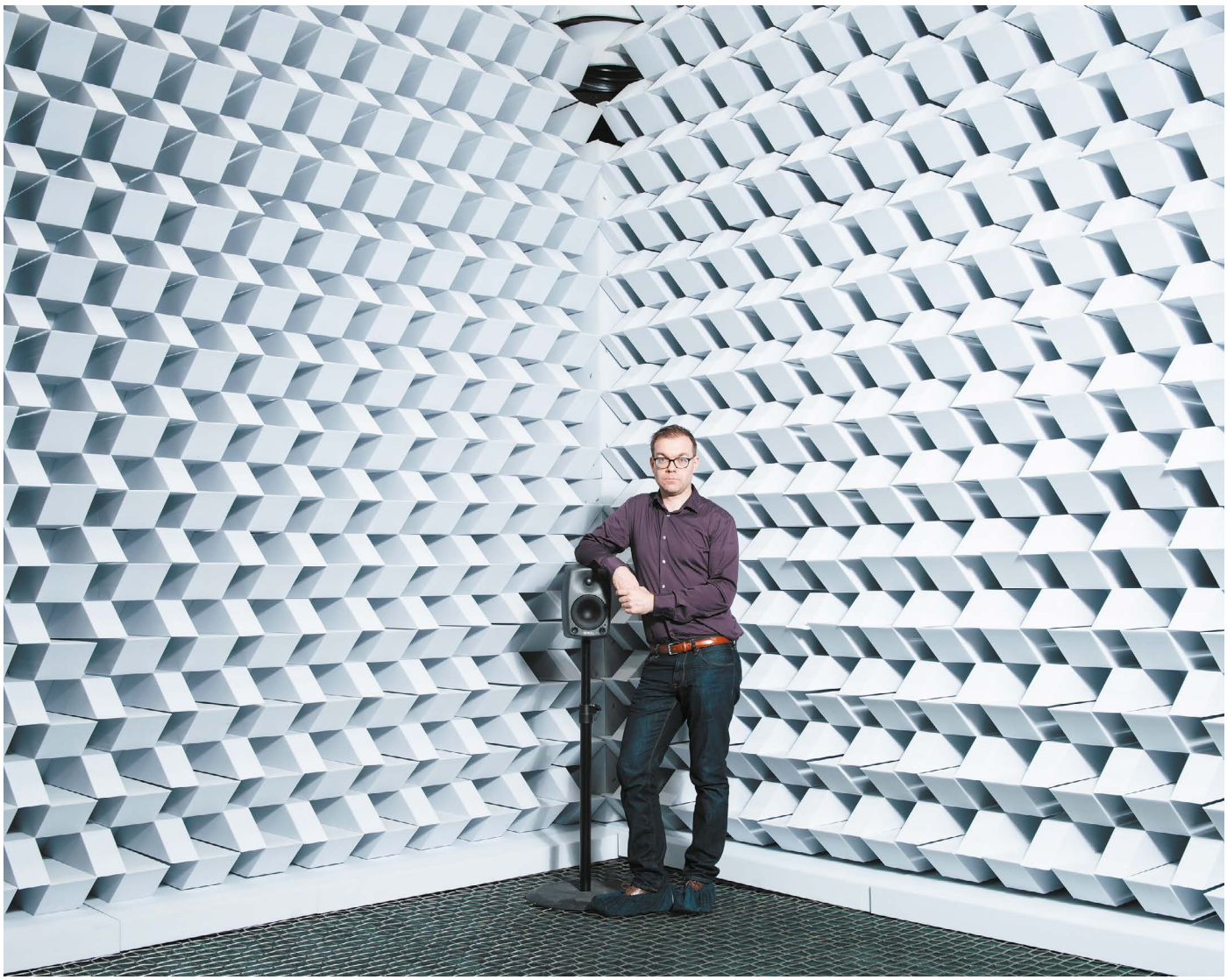
She takes care of him during the evening hours. So, I have from 10 p.m. to 2 a.m. to focus on duties such as writing papers or preparing lectures. During the day, I take care of my son so that my wife can have at least four hours of focus. We have expanded the capacity of Skyroom, the university's virtual class platform, which was produced by an Iranian start-up company. For the past month or so, my wife and I have nearly managed to synchronize my son's naps with my class schedule, although we are not always successful. If he wakes up during the class, my wife handles him. Outside of that class hour, I take care of him.

My appointments with postdocs and students are set for approximate times and are occasionally rescheduled to accommodate my wife's working hours.

Because my daughter likes to watch her favourite cartoon in the living room and my son is asleep in the bedroom while I teach my afternoon class, I usually hold my online undergraduate course while sitting on the balcony of our apartment.

Seyed Akbar Jafari is a condensed-matter physicist at Sharif University of Technology in Tehran.

Interviews by Robin Lloyd.
Interviews have been edited for length and clarity.



Where I work
Jukka Pätynen

Photographed for *Nature* by Jarkko Mikkonen.

As an acoustician with the Helsinki-based consultancy Akukon, I design and test the sound characteristics of various spaces, from performance venues to research laboratories.

This year, I was part of a project to completely renovate the acoustic labs of Aalto University in Espoo, Finland. The room in this picture is an anechoic space: it's designed so that no surface reflects sound. Hard foam wedges cover the walls to absorb any incident sound; the room itself is a box within a box, with the inner space floating on elastic materials for vibration isolation.

On the decibel scale, 0 dB is the lower limit of human hearing. After the renovation, we measured the background sound level in this space to be −2 dB — the limit of our measurement devices — but our calculations suggest that the actual background noise could be as quiet as −10 dB.

When I was a postdoc at Aalto, we reproduced in these labs the acoustic characteristics that we had measured in concert halls across central Europe. We could 'transport' listeners in the lab between various halls while keeping everything else constant, from the musical performance to the mood of the listener. It enabled us to directly compare the spaces and identify the main ingredients of good room acoustics.

Some of the acoustic designs that I have worked on at Akukon derive from this project. We helped to design a club-type music venue, G Livelab in Helsinki, in which sound engineers use a virtual acoustic system to simulate different halls. The timbre of the sound and how big the space seems to be can be pretty freely adjusted, so the room suits everything from rock bands to string quartets.

Right now I'm working mainly from home, but I can still make site visits to take acoustic measurements. When society eventually reopens, I will be interested to see whether the general appreciation of concerts and similar gatherings has grown over this extraordinary period.

Jukka Pätynen is an acoustician at Akukon in Helsinki, Finland.
Interview by James Mitchell Crow.